UNSW-NB15 dataset feature selection and network intrusion detection using deep learning

Article · January 2019

CITATIONS
41

READS
5,128

2 authors:

V. Kanimozhi
Sathyabama Institute of Science and Technology
7 PUBLICATIONS **344** CITATIONS
SEE PROFILE

Prem Jacob
Sathyabama Institute of Science and Technology
91 PUBLICATIONS **1,072** CITATIONS
SEE PROFILE

UNSW-NB15 Dataset Feature Selection and Network Intrusion Detection using Deep Learning

V. Kanimozhi, Prem Jacob

Abstract*: Anomaly detection system in network, monitors and detects intrusions in the networking area, which is referred to as NIDS, the Intrusion Detection System in Networks. There are numerous network datasets available in networking communications with relevant and irrelevant features drastically decreases the rate of intrusion detection and increases False Alarm Rate. The benchmark network dataset available is UNSW-NB15 dataset was created in 2015. The top significant features are proposed as feature selection for dimensionality reduction in order to obtain more accuracy in attack detection and to decrease False Alarm Rate. We apply a combination fusion of Random Forest Algorithm with Decision Tree Classifier using Anaconda3 (free and open-source distribution of Python3) and package management system Conda in which 45 features have been decreased to the strongest four features. The proposed system detects normal and attacks with a better accuracy using Deep Learning technique.

Index Terms : data visualization; feature selection; intrusion detection; Artificial Neural Network; UNSW-NB15 Dataset.

I. INTRODUCTION

The Australian Center for Cyber Security (ACCS), Cyber Range Lab has made an IXIA PerfectStorm tool for creating the hybrid of synthetic contemporary attack behavior and real modern activities. To find the 100 GB of raw traffic (for example, Pcap files), a Tcpdump tool is used. By utilizing the tools like Argus and Bro-IDS, and by developing 12 algorithms collectively 45 features with the class label of total 2,540,044 records. Dataset Features are broadly categorized into 6 subsets as Labeled Features, Time Features, Content Features, Flow Features, Basic and features which are additionally generated. More attacks on UNSW-NB15 are further categorized as 9 various types, namely Worms, Fuzzers, DoS, Exploit, Reconnaissance, Backdoor, Analysis, Shellcode, and Generic[6]. By further categorizing Additional Generated Features, two sub-groups formed namely Connection and General Purpose Features. The numbering of features from 36-40 and from 41-47 are called as General Purpose and Connection Features respectively. But now the revised dataset of UNSW-NB15 dataset consists of only 45 features.

II. RELATED WORK

The first research carried out by Moustafa and Slay using algorithm Association Rule Mining(ARM) and listed important features for every attack type of higher frequencies in detecting intrusions.

Figure 1. Frequented features in ARM algorithm

According to Janarthanan, Tharmini and Zargari, Shahrzad, few of the Attribute Selection methods are applied to UNSW-NB15 Dataset like GreedyStepwise, InfoGainAttributeEval, CfsSubsetEval (all 3 attribute evaluator) and Ranker. In weka Environment, the Random Forest and some Machine-learning algorithms are enforced to this dataset by examining suggested features.

Table 1. Machine Learning Techniques suggested features are

service	sbytes	sttl	smean	ct_dst_sport_ltm

III. METHODOLOGY

3.1 Total Features, attack categories and correlation matrix in UNSW-NB15 Dataset

This dataset consists of forty-five attributes or features and the strongest attributes can be proposed to detect more accuracy. The dataset contains some irrelevant and redundant features, which is unimportant. Feature Selection plays an important role in achieving more accuracy in intrusion detection. By including records of normal traffic and all attack types, UNSW-NB15 dataset is broadly classified into Testing and Training datasets with #175, 341 and #82, 332 records respectively. These two datasets include 45 features (see Table 2) for further reference. Please note that the features Itime, stime, sport, scrip, and dstip are missing in both datasets (Testing and Training) [14] and in the whole UNSW-NB15 dataset, id is the first feature is not listed.

Revised Version Manuscript Received on January 25, 2019.

V. Kanimozhi, School of Computing,, Sathyabama Institute of Science & Technology,Chennai, India

Dr. Prem Jacob, School of Computing,, Sathyabama Institute of Science &Technology,Chennai, India

The 45 features are represented in the below Table 2.

Attribute Number	Attribute Name	Attribute Number	Attribute Name
1	id	23	dtcpb
2	dur	24	dwin
3	proto	25	tcprtt
4	service	26	synack
5	state	27	ackdat
6	spkts	28	smean
7	dpkts	29	dmean
8	sbytes	30	trans_depth
9	dbytes	31	response_body_len
10	rate	32	ct_srv_src
11	sttl	33	ct_state_ttl
12	dttl	34	ct_dst_ltm
13	sload	35	ct_src_dport_ltm
14	dload	36	ct_dst_sport_ltm
15	sloss	37	ct_dst_src_ltm
16	dloss	38	is_ftp_login
17	sinpkt	39	ct_ftp_cmd
18	dinpkt	40	ct_flw_http_mthd
19	sjit	41	ct_src_ltm
20	djit	42	ct_srv_dst
21	swin	43	is_sm_ips_ports
22	stcpb	44	attack_cat
		45	label

Table 2. UNSW-NB15 dataset listed Features [14]

The 9 types of attack categories are namely Analysis, Fuzzers, Exploits, Shellcode, Reconnaissance, DOS, Backdoors, Shellcode, and Worms of UNSW-NB15 Training Dataset and as represented by the graph in (Figure 2).

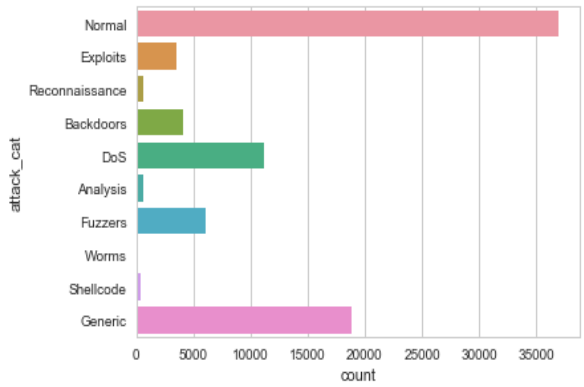


Figure 2. Attack categories (9 types) in UNSW-NB15 Training Dataset

The Correlation matrix of features in UNSW-NB15 dataset is shown below, where ID feature is not included. This graph in (Figure 3) makes us understand the correlation relationship between features in the dataset. For e.g., features sttl is closely correlated to feature dttl, feature sinpkt is closely correlated to dinpkt. Likewise, the Feature Importance can be understood using this Correlation Matrix Graph.

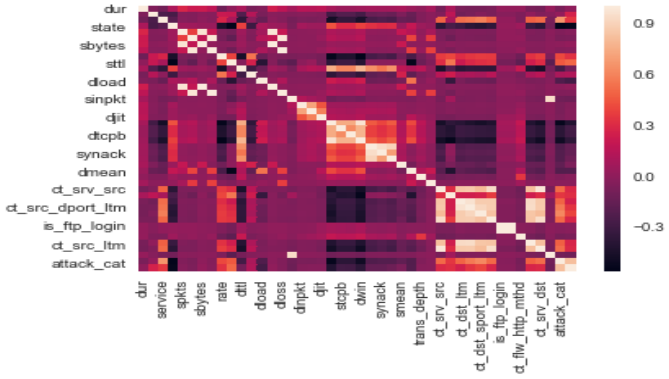


Figure 3. Correlation Matrix of UNSW-NB15 Dataset

3.2 Feature Importances

The proposed study in order to avoid irrelevant features and to decrease the dimensionality of the dataset it has been used Anaconda 3 with Python 3 distribution and Jupyter notebook using the Python Machine learning library sklearnkit. The proposed algorithm is RandomForestClassifier, which is an ensemble classification algorithm. A group of classifiers is called as Ensemble classifier. This Ensemble classifier gives the advantage of predicting the target using multiple classifiers than using a single classifier. The randomly created decision trees are called as classifiers. The target prediction is based on the majority of votes, considering each decision tree as a single classifier. The maximum number of votes received by the target class is considered as the final predicted target class. It provides maximum accuracy with the combination fusion of recursive feature elimination from sklearn feature selection method. The top four features are sttl, ct_dst_src_ltm, sbytes, sload, provides 98.3 percent accuracy. Combination fusion of Decisiontree and RandomForestClassifier which provides 98.3 percent has listed the best four features are as sbytes, sttl, sload, ct_dst_src_ltm. The graphical representation of Feature Importances and the top four features are labeled in the graph as shown below in (Figure 4).

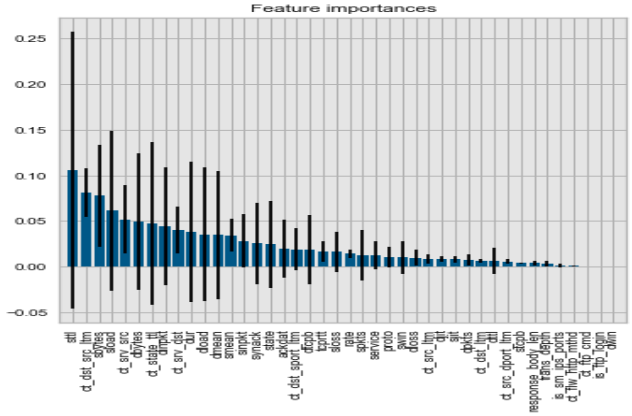


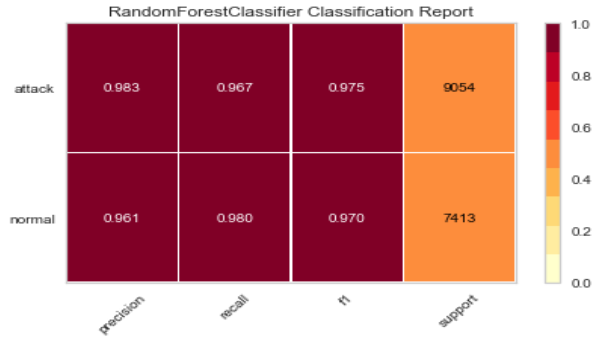
Figure 4. Feature Importances of UNSW-NB15 Dataset

3.3 The RandomForestClassifier Classification Report

The proposed RandomForestAlgorithm use Yellowbrick's classification report, a model visualizer the displays precision, recall, F1 scores, and support. This tool integrates numerical scores as well as color-coded heatmaps in order to support easy interpretation and detection

.

Table 3. RandomForestClassifier Classification Report



In the above table, Precision is the proportion of right records to the total right records. The proposed precision of attack in the training dataset is 98.3 percent.

A recall is the proportion of right records to the accurate number of correct actual records and the proposed recall is 98 percent

A measure of accuracy test is known as the F1 score and formulated as an average weight of precision and recall, the proposed F1 score of attack is 97.5 percent.

Also proposed Feature selection Algorithm using Recursive Feature Elimination (RFE) as its title suggests, recursively removes features and builds a model using the best attributes and calculates accuracy. It imports RFE from sklearn.feature_selection and pass to RandomForestClassifier with the best top ranking features to select. The best features are sbytes, sttl, sload, ct_dst_src_ltm which produces 98 percent accuracy.

V. ARTIFICIAL NEURAL NETWORKS

Artificial Intelligence is the inspiration from the human mind. It is implemented in the proposed system as Artificial Neural Networks also abbreviated as ANN. The angle of ANN is to find out from the model instances, analyzes the data from the learned instances by that it's capable of classifying multi-classes with higher accuracy. During this deep learning, learning refers to modify the data by providing weights and connect between the nodes of a layer in neural networks and also the wit is set by the formula and also with the kind of artificial networks selected for coaching use. If we've got uncountable records, Deep Learning is that the better option in real-world information and it outperforms the performance of the opposite classifier models. MultiLayerPerceptron (MLP) is that the planned system, that could be a Feedforward Multilayer Neural Networks that gives high machine power by combining several straightforward units referred to as neurons and verify the edge to perform parallel tasks. It classifies well the traditional and attack instances in this dataset which produces a lot of accuracy within the planned system.

V. RESULTS

5.1 Creating an Artificial Neural Network with Anaconda, Jupyter Notebook, and SciKit- Learn

To build this Artificial Neural Network, we use Anaconda 3.0 and the latest Scikit version 0.19.1 and Pandas version 0.23.1 in Jupyter Notebook. It can be installed through pip or Miniconda (Package Manager of Anaconda).

5.2 Receiver operational Characteristics Curve (ROC)

ROC curve is employed to visualize the performance of multi-dimensional classification data. It is being considered as one of the distinguished evaluation metrics for evaluating any classification model's accuracy.

Let's model the neural net and do prediction process.

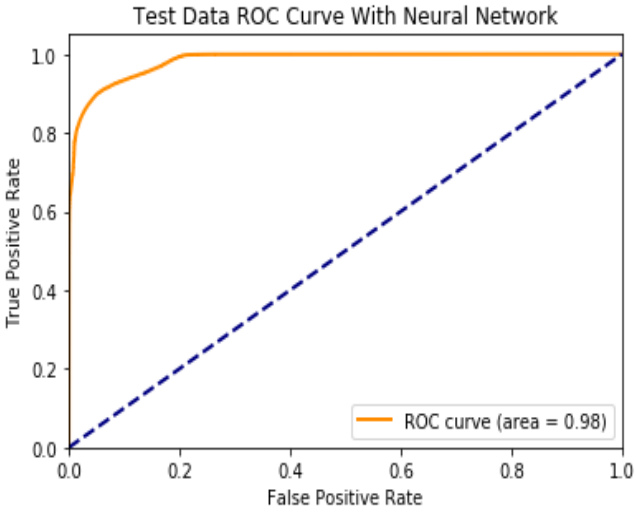


Figure.5 ROC CURVE

The classifier model runs on UNSW-NB15 with the strongest features to produce ROC curve area = 0.98 in Fig.4

5.3 Classification Report of the proposed system

Training Data Performance Metrics					
	Accuracy	Precision	Recall	F1	AUC
0	0.96	0.97	0.96	0.97	0.99
Test Data Performance Metrics					
	Accuracy	Precision	Recall	F1	AUC
0	0.89	0.99	0.85	0.91	0.98

Figure.6. Classification Report of ANN

5.4 Confusion Matrix

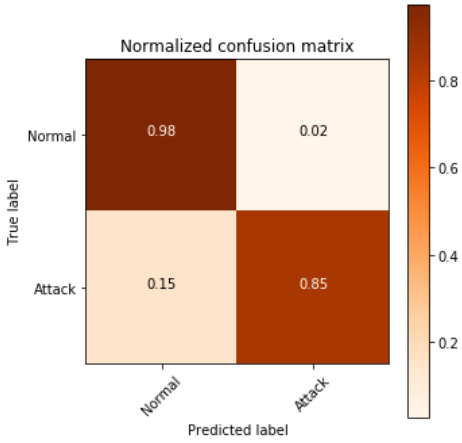


Figure.7 Confusion Matrix

It provides insights of the quantity of positive and negative predictions and conjointly summarizes the count of traditional and malicious attacks during this model and therefore the below graph is shown with samples however higher it identifies the conventional and malicious attack. therefore overall confusion Matrix outperforms the analysis metrics of this model.

VI. CONCLUSION

This proposed system now has only top 4 features of UNSW-NB15 Dataset that is extracted from the total 45 features by using combination fusion algorithm of RandomForest algorithm and Decision tree classifier by Research work goes on network intrusion detection within the realistic cyber dataset to seek out zero-day attacks and existing attacks with a lot of accuracy and to attain higher performance using Anaconda3 (an open source distribution of Python3).

REFERENCES

1. "Network Intrusion Detection and Prevention: Concepts and Techniques ", by Ghorbani A., Lu W., and Tavallaee M., 2010, Springer Science, LLC.
2. "Feature selection and intrusion classification in NSL-KDD cup 99 datasets employing SVMs", by Pervez M. S. and Farid D. M. The 8th International Conference on Software, Knowledge, Information Management and Applications (SKIMA 2014), Dhaka, 2014, pp.1-6.
3. "Unsw-nb15: A comprehensive data set for network intrusion detection," in MilCIS-IEEE Stream, Military Communications and Information Systems Conference by Moustafa N. and Slay J., Canberra, Australia, IEEE publication,2015
4. The significant features of the UNSW-NB15 and the KDD99 sets for Network Intrusion Detection Systems", by Moustafa N. and Slay J., the 4th International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS 2015), collocated with RAID 2015, 2016, [Online]available:
5. "Feature selection and intrusion classification in NSL-KDD cup 99 datasets employing SVMs,"by Pervez M. S. and Farid D. M., The 8th International Conference on Software, Knowledge, Information Management and Applications (SKIMA 2014), Dhaka, 2014, pp.1-6.
6. An investigation into discrepancies in findings with the KDDCUP'99 data set and multi-objective evolution of neural network classifier ensembles from imbalanced data, by Engen, Vegard. Machine learning for network-based intrusion detection: Diss. Bournemouth University,2010.
7. "Robust Preprocessing and Random Forests Technique for Network Probe Anomaly Detection.," by Kumar, G. Sunil, and C. V. K. Sirisha, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-1, Issue-6, January 2012.
8. "Improving the Intrusion Detection using Discriminative Machine Learning Approach and Improve the Time Complexity by Data Mining Feature Selection Methods",by Bajaj and Arora, International Journal of Computer Applications (0975-8887), Volume 76-No.1, August 2013. [Onine] available: http://research.ijcaonline.org/volume76/number1/pxc3890587.pdf
9. "Toward Generating a New Intrusion Detection Dataset and intrusion Traffic Characterization",by Iman Sharafaldin, ArashHabibiLashkari, and Ali A. Ghorbani, 4th International Conference on Information Systems Security and Privacy (ICISSP), Portugal, January 2018.
10. "Intelligent intrusion detection system using artificial neural networks," by Alex Shenfield, David Day, and Aladdin Ayesh, vol. 4, no.2, pp. 95-99, June 2018.
11. Feature Selection in UNSW-NB15 and KDDCUP'99 datasets by JANARTHANAN, Tharmini and ZARGARI, Shahrzad (2017) In: 2017 IEEE 26th International Symposium on Industrial Electronics (ISIE),. IEEE.
12. "Intrusion detection system: A comprehensive review"J.Netw. Comput. Appl., Rev., 36 (1), pp. 16-24,2013 by Liao H.-J., Lin C.-H.R., Lin Y.-C., and Tung K.-Y. [Online]. Available https://www.kdnuggets.com/2016/10/beginners-guide-neural-networks-python-scikit-learn.html
13. "The trends of intrusion prevention system network, in: 2010" by D. Stiawan, A.H. Abdullah, and M.Y. Idris, 2nd International Conference on Education Technology and Computer, vol. 4, pp. 217-221, June 2010.
14. "Network intrusion detection based on a general regression neural network optimized by an /5improved artificial immune algorithm."Rev.,10 (3), 2015 by Wu J., Peng D., Li Z., Zhao L., and Ling H. Wu J., Peng D., [Online] Available https://www.ncbi.nlm.nih.gov/pubmed/25807466.
15. "Neural networks for classification: A survey" by Zhang G.P. IEEE Trans. Syst. Man Cybern. C, Rev., 30 (4), pp. 451-462, 2000.