Accepted Manuscript

Artificial Intelligence based Network Intrusion Detection with hyper-parameter optimization tuning on the realistic cyber dataset CSE-CIC-IDS2018 using cloud computing

V. Kanimozhi, T. Prem Jacob

PII: S2405-9595(18)30597-6
DOI: https://doi.org/10.1016/j.icte.2019.03.003
Reference: ICTE 197

To appear in: *ICT Express*

Received date : 13 December 2018
Accepted date : 5 March 2019

Please cite this article as: V. Kanimozhi and T.P. Jacob, Artificial Intelligence based Network Intrusion Detection with hyper-parameter optimization tuning on the realistic cyber dataset CSE-CIC-IDS2018 using cloud computing, *ICT Express* (2019), https://doi.org/10.1016/j.icte.2019.03.003

Artificial Intelligence based Network Intrusion Detection with hyper-parameter optimization tuning on the realistic cyber dataset CSE-CIC-IDS2018 using cloud computing

V. Kanimozhi , Dr. T. Prem Jacob

Research Scholar, Department of CSE,
Sathyabama Institute of science and technology, Chennai, India
Email : kanimv@yahoo.co.in

Associate Professor, Department of CSE,
Sathyabama Institute of science and technology, Chennai, India
Email : premjac@yahoo.com

Abstract

One of the latest emerging technologies is artificial intelligence, which makes the machine mimic human behaviour. The most important component used to detect cyber attacks or malicious activities is the intrusion detection system (IDS). Artificial intelligence plays a vital role in detecting intrusions and widely considered as the better way in adapting and building IDS. In modern days, neural network algorithms are emerging as a new artificial intelligence technique that can be applied to real-time problems. The proposed system is to detect a classification of botnet attack which poses a serious threat to financial sectors and banking services. The proposed system is created by applying artificial intelligence on a realistic cyber defence dataset (CSE-CIC-IDS2018), the latest IDS Dataset in 2018 by Canadian Institute for Cybersecurity (CIC) on AWS (Amazon Web Services).

The proposed system of Artificial Neural Networks provides an outstanding performance of Accuracy score is 99.97% and an average area under ROC(Receiver Operator Characteristic) curve is 0.999 and an average False Positive rate is a mere value of 0.03. The proposed system of Artificial Intelligence-based Intrusion detection of botnet attack classification is powerful, more accurate and precise. The novel proposed system can be applied to conventional network traffic analysis, cyber-physical system traffic analysis and also can be applied to the real-time network traffic data analysis.

Keywords : Artificial Intelligence, AWS, CSE-CIC-IDS2018, hyper-parameter optimization, realistic network traffic cyber dataset

1. INTRODUCTION

The objective of network intrusion detection is identifying and monitoring malicious activities. Most of the current IDSs canbe divided into two main categories. They are signature-based and anomaly-based IDS. In signature-based IDS, by comparing the already known attacks with the incoming network traffic tries to detect the intrusions, that are stored in the database as signatures. Existing attacks are well detected by IDS, but it often fails to detect novel (unseen) attacks. The next category is called anomaly-based IDS. The normal traffic is modelled by the IDS models through learning patterns in the training phase. The deviations from these learned patterns are labelled as anomaly or intrusion. The implementation of real-time anomaly-based IDS is a herculean task because of the rapid increase in the network traffic behaviour and very limited availability of computational resources (computation time and memory).

There is another challenge and that is the risk of overfitting due to the high dimensional feature space and the model complexity of IDS. Artificial intelligence (AI) based techniques play a crucial role in the development of IDS and has more advantages over other techniques. There is no appropriate and well-defined techniques to solve the anomaly detection problems. The proposed system will help the better understanding of different directions in which research has been done in the field of IDS. They are beneficial for those who are interested in applications of AI-based techniques to IDS and related fields. In this paper, we proposed an experimental approach of Artificial Neural Networks with hyper-parameter optimization on the realistic new IDS cyber dataset (cse-cic-ids2018) included most of the up-to-date attacks (PCAP) along with labelled flows covering more than 80 features (CSV) which obtained through cloud computing on AWS services for intrusion detection in order to provide more accurate accuracy.

2. BACKGROUND AND RELATED WORK

A lot of research work carried out in Network Intrusion Detection System either in Host-Based Intrusion Detection (HIDS) or Network-based intrusion detection (NIDS) and Artificial Intelligence, but there is no comprehensive reliable cyber dataset which covers both contemporary and modern-day attacks for network intrusion detection system. According to Alex Shenfield and his co-authors stated that the research carried out an offline approach for detecting shellcode patterns within data [2].

Networks are more vulnerable day by day due to modern attacks. In this proposed system, we make use of realistic latest cyber dataset which compromises of both existing

attacks and zero-day attacks by Canadian Cybersecurity obtained through cloud computing.

3. METHODOLOGY

3.1 BOTNET

A botnet is an attack which is coined from two words "Robot" and "Network". It is a network which can be operated or commanded by remotely controlled computers. Moreover, it is nothing but a malware that makes the system or server to be controlled and commanded remotely by an operator.

Cyber apocalypse - It is another malware and poses a serious threat to networks which produces the impact of an army of bots. E.g. of bots. Zeus,Ares etc. Denial-of-service attacks are launched by Botnets of zombie computers which can be propagated through Drive-by-downloads and spam emails.

It performs various criminal and malicious activities like stealing information especially in the banking and financial sectors by logging and grabbing customers information. It hijacks the confidential information like username, password, and other sensitive information.

Crypto-Locker ransomware - It is a luxurious software which attacks the window operating system by encrypting all the files in the user system with RSA-2048 public key. It claims hefty ransom in order to decrypt the file. It is astounding fact that the virus earned $30 million in hundred days.

Credential stuffing is a malicious activity which handles the automated injection attack by making usage of botnets in it to access the online services by stealing the significant credentials. Researchers from Akamai reported the fact that 30 billion malicious login attempts were made between November 2017 to June 2018 from the states of the United States, Russia, and Vietnam.

3.2 Artificial Neural Networks

Artificial Neural Networks are a machine learning framework that attempts to mimic the learning pattern of natural biological neural networks. Biological neural networks work in the sense that the dendrites receive inputs which are said to be presented in the interconnected neurons of the human brain. Based on these inputs, through an axon to another neuron, they produce an output signal. We will try to mimic this process using Artificial Neural Networks (ANN), just refer to as neural networks from now on. Neural networks are the foundation of deep learning. It is a subset of machine learning responsible for some of the most exciting technological advances today!

3.3 Multi-layer Perceptron tuning with hyper-parameter optimization Classifier Model

A Perceptron has the following one or more inputs, a bias, an activation function, and a single output. The Perceptron gets inputs, applies some weight, and the output is produced by the activation unit which receives the weighted inputs. The neural network can be modeled by adding perceptrons layer together to form Multi-layer perceptrons of an Artificial Neural Network. You'll have an input layer that directly takes in your data and an output layer, which will create the resulting outputs. Any layers in between are known as hidden layers. This is because they don't directly "see" the feature inputs within the data you feed in or the outputs.

We will use Multi-layer Perceptron (MLP) that implements a multi-layer perceptron algorithm. To specify the number of layers, as well as the number of neurons for each layer, is enabled by having multiple hidden layers. Multi-layer Perceptron is sensitive to feature scaling. Therefore, Scaling your data should be advisable.

As for hyper-parameter optimization, GridSearchCV Optimization technique is used. Tuning a neural network for optimization is a herculean task and it is a lengthy process. It operates on parallel and can be iterated, with 10-fold cross-validation. We model our neural network by starting with two layers [2].

Solver has been picked in this model as 'lbfgs'. And try to findalpha parameter using L2 regularization. Better prediction and accuracy will not be generated without regularization techique. In this model, we classify the intrusion detection ("Benign" or "Malicious") based on the output.

Best F1 Score is: 0.9991678456370812
Best Parameter is: {'alpha': 1e-05, 'hidden_layer_sizes': (9, 4)}

4. Implementation

4.1 CSE-CIC-IDS2018

We built an MLP Classifier model on realistic cyber defence dataset by Canadian Institute for Cybersecurity (CIC) on AWS (Amazon Web Services). *Datasets* by *CIC* and ISCX are used around the world for security testing and malware prevention. Knowledge on AWS is a must for accessing that dataset which is stored in Resource type **-S3 Bucket** and Amazon Resource Name(ARN) is **arn:aws:s3:::cse-cic-ids2018**and also **AWS Region C**a-central-1 under License.

License: http://www.unb.ca/cic/datasets/ids-2018.html

It includes a detailed description of intrusions along with abstract distribution models for the following: applications, protocols, or lower level network entities. The final dataset includes seven different attack scenarios: Brute-force, Heartbleed, Botnet, DoS, DDoS, Web attacks, and Infiltration of the network from inside. The attacking infrastructure includes 50 machines. The victim organization has 5 departments and includes 420 machines and 30 servers. The dataset includes the captures network traffic and system logs of each machine, along with 80 features extracted from the captured traffic using CICFlowMeter-V3 [1].

4.2 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).

4.3 Receiver Operating Characteristics Curve

ROC (Receiver Operating Characteristics**)** curve is used to visualize the performance of multi-dimensional classification data. It is being considered as one of the most prominent evaluation metrics for evaluating any classification model's a ccuracy. It is also referred to as AUROC (Area Under the Receiver Operating Characteristics).

Let's model the neural net and do prediction.

To get the whole evaluation metrics, I have created tw functions. The calculate_auc function also produces ROC. To create a data frame for the easy summary of performance metrics, and that has been done by pandas.

5. RESULTS

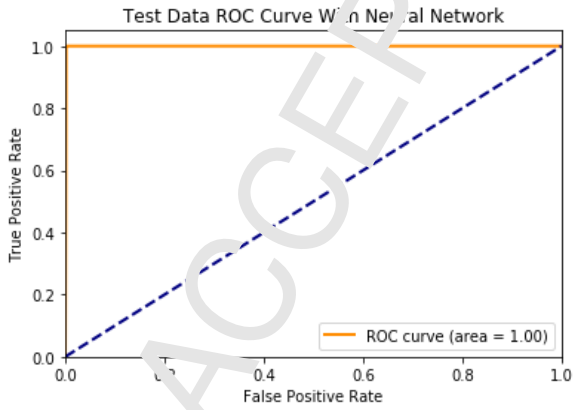

Fig 1. ROC CURVE

5.1 ROC Curve

The curve generated in Fig.1 when True positive versus against False Negative rate at various threshold points and the curve implies how well the binary classifier discriminated bet ween twodifferent classes i.e., Benign or malicious. The classif ier model runs a sample of 1048575 records with 80 features a nd optimize it with 10 Fold Cross Validation to produce the R OC curve in Fig.1.

5.2 AUC SCORE

It is the area under the ROC Curve, and it summarizes the perf ormance of the binary classifier.Higher the score, better the classifier model performance

AUC SCORE : 0.99[illegible]0

5.3 Confusion Matrix

It gives insights of the number of positive and negative predi ctions and also summarizes the count of normal and malicious attacks in this model and the below graph isshown with samples how 100% it identifies the normal and malicious botnet attack. So overall confusion Matrix outperforms the eva luation metrics of this model.

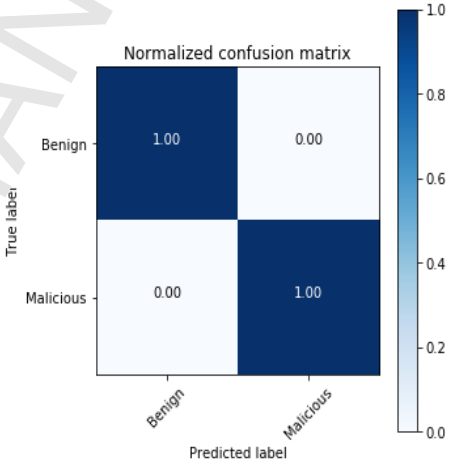

Fig.2 Confusion Matrix of Neural Network

5.4 Classification Report of Neural Network Model

Table 1. Classification Report of Neural Network

Training Data Performance Metrics					
Accuracy	Precision	Recall	F1	AUC	
0	1.0	1.0	1.0	1.0	1.0
Test Data Performance Metrics					
Accuracy	Precision	Recall	F1	AUC	
0	1.0	1.0	1.0	1.0	1.0

Neural Network Model Train Accuracy: 1.0

Neural Network Model Test Accuracy: 0.99975

5.5 Default MLP Classifier Model Comparison

If the model has not been set by any parameter, the default alpha is 0.0001 and hidden_layer_sizes is 100 neurons in a single layer. The default model is over-fitting. This happens a lot for neural networks. You can visualize the accuracy score and the power of parameter optimization can be realized.

Neural Network Model Train Accuracy: 0.99983
Neural Network Model Test Accuracy: 0.9995

6. CONCLUSION

The proposed system can be extended to detect remaining classes of attacks in this realistic dataset which includes all real-time and existing attacks. The framework used in this artificial Intelligence Scikit learn framework optimization is based on CPU (Central Processing Unit) not on GPU (Graphics Processing Unit), the optimization can be powerfully tuned by other such frameworks like Google's open sourced TensorFlow. The performance issue is a common task when we come across pandas to work with larger data (100 gigabytes to multiple terabytes), but Spark is an open-sourced Apache Framework used for big data processing can handle parallel computing with massiv datasets, ranging from 100 gigabytes to multiple terabytes across clustered computers.

Conflict of interest

I, declare there is no conflict of interest in thi paper.

References

[1] Iman Sharafaldin, ArashHabibiLashkari, and Ali A. Gh bani, "Toward Generating a New Intrusion Detection D t andIntrusion Traffic Characterization", 4th International Confe nce o Information Systems Security and Privacy (ICISSP), Portugal, J ary 2018.

[2] Alex Shenfield, David Day, and A addin Ayes . "Intelligent intrusion detection system usingartificial neur netw rks," vol. 4, no.2, pp. 95-99, June 2018.

[3] D. Stiawan, A.H. Abdullah and M.Y. Id s, "The trends of intrusion prevention system network, i 2010" 2 d International Conference on Education Technology and Co puter, vol. 4 pp. 217-221, June 2010.

[4] Singh R., Kumar H., Si gla R.K., andKetti R.R. "Internet attacks and intrusion detection sys m: A re iew of the literature"Online Inform. Rev., 41 (2), pp. 171-18 . 2017.CrossRefView Record in ScopusGoogle Scholar.

[5] Liao H.-J., Lin -H. Lin Y.-C., and Tung K.-Y. "Intrusion detection system: A comp hen ive review"J.Netw. Comput. Appl., Rev., 36 (1), pp. 16-24,2013. [Online]. Available https://www.kdnuggets.co /2016/10/beginners-guide-neural-networks-python-scikit-learn.html.

[6] Zhang G.P. "Neural networks for classification: A survey" IEEE Trans. Syst. Man Cybern. C, Rev., 30 (4), pp. 451-462, 2000.

[7] Wu J., Peng D., Li Z., Zhao L., and Ling H. "Network intrusion detection based on a general regression neural network optimized by an /5improved artificial immune algorithm."Rev.,10 (3), 2015. [Online] Available https://www.ncbi.nlm.nih.gov/pubmed/25807466.

[8] Rosenblatt F. "The perceptron: A pro abilistic model for information storage and organization in the brain" P yc l Rev., 65 (6), pp. 386-408, 1958.

[9] Data Science , Machine Learni g blog . [Online] Available https://www.mydatahack.com/predic rnet-popularity-by-optimising-neural-networks-with-r/.

AUTHORS PROFILE

I am kanimozhi, worked s an Assistant Professor(Computer Science) for th e years and also working as a Software Trainer (C, C+ , Java, ython, Python Libraries(Numpy, Scipy, Pandas), M chin Learning, Python Implementation Jupyter N tebook Currently pursuing Ph.D(CSE) in Big Data Anal tic of La ge Scale Network Based security using Python Anac nda 3.0 & Spark Implementation in Sathyabama I stitute of Science and Technology, Chennai.

Dr. T. Prem Jacob received the B.E degree in Computer Science and Engineering from C.S.I Institute of Technology, Manonmaniam Sundaranar University, Nagercoil, India in 2004, M.E degree in Computer Science and Engineering from Sathyabama University, Chennai, India in 2006 and Ph.D. degree from Sathyabama University, Chennai, India. He is an Associate Professor of Computer Science and Engineering in Sathyabama Institute of Science and Technology Chennai. He has participated and presented many Research Papers in International and National Conferences. His area of interests includes Software Engineering, Data mining and Data warehouse and Cloud computing.

CONFLICT OF INTEREST

I, V. Kanimozhi declare that there is no Conflict of Interest in this paper.