Performance Analysis of Intrusion Detection System using Deep Learning Techniques

Usnish Mukherjee
(19MA60R11)

10.11.2020

Contents:

Objective

- Hardwares or softwares that are used for monitoring and analyzing data flow between the hosts in a network to detect security threats are known as **Intrusion Detection System(IDS)**.
- To predict and prevent the intrusion attacks, deep learning techniques are applied here on big data to come up with a model that can differentiate between normal and attacked network traffic flow.

Introduction

- The expanding nature of the use of the Internet and the rapidly growing nature of the volume of data are giving more opportunity to the hackers to initiate their harmful attacks.
- On the contrary, researchers and developers aim to raise the efficiency of early harmful attack forecasting and detection.

- Intrusion detection systems in general use two known methods to analyze flow and detect attacks:
  - Signature based detection.
  - Anomaly detection.

Cont.

- With the expansion of connectivity and networked systems, the network traffic flow is currently considered as big data.
- Big data has 7 'V' properties, that are Volume, Variety, Velocity, Variability, Veracity, Visualization and Value.
- In this project mainly the Volume property of big data is utilised since the size of the data is nearly 7GB containing around 17 million records with more than 75 features. Also the Variety property came into help to some extent.
- Though, in most of the systems firewalls are used to avert the cyber attacks, Intrusion detection systems (IDSs) play a notable role to enhance the system security.

Literature Review

The summary of the literature review is summarized below

Paper	Dataset	Method	Accuracy
Osama Faker and Erdogan Dogdu [1]	CICIDS-2017 UNSW NB15	GBT(Binary) , DNN-4(Multi) DNN-4(Both)	99.9%,99.56% 99.16%,97.01%
X. Zhang and J. Chen [2]	KDDCup-99	RBM-DBN	97.16%
G. C. Fernández et al. [4]	ISCXIDS-2012 CICIDS-2017	DNN(with and without IP)	96.77%
Rahul Vigneswaran K et al. [5]	KDDCup-99	DNN-3	93%

Table 1 - Summary of literature review

Proposed Work

- The method proposed in the work is to classify between normal and attacked traffic flow obtained from the **CSE-CIC-IDS2018** dataset.

- Previously due to lower number of attack types and without the availability of benchmark datasets machine learning techniques were used. But with the rapidly increasing size of network data ,deep learning techniques are inevitable to increase efficiency, accuracy and response time of IDS.

- In this work a deep neural network technique is used with 4 hidden layers along with one input and one output layer. With the large number of records along with a large number of features the deep learning technique must be more useful than traditional machine learning techniques. In order to show that it is compared with Random Forest.

Concepts used

- **Big Data** - Here the 'Volume' property of big data is used. To simulate the real world scenario of network traffic, the size of the dataset is very large (nearly 7GB) with nearly 17 million records with more than 75 attributes. Without the help of deep learning this amount of large data cannot be handled properly. The 'Variety' property also came to help.

- **Deep Learning** - Deep learning is a specialized machine learning technique which is basically a neural network. Deep learning algorithms operate using a number of levels/layers. The layers are interconnected and the output of the previous layer is fed as input to the next level,so it is called feed-forward network.

Cont...

- **Backpropagation** - Backpropagation is a method to train neural networks. Based on the loss in the previous epoch the weights of the connections and biases of the neurons are modified based on some rule. The algorithm propagates backward from the last layer to avoid unnecessary calculations and is an example of dynamic programming.

- **Random forest** - Random forest classifier is an ensemble classifier made up of multiple decision trees. Each decision tree is trained on different data samples and the final class is decided by majority election/voting.

Model Used

- The deep neural network used here is made up of 4 hidden layers between 1 input and 1 output layer which are fully connected with each other. ReLU function is used in hidden layers.
- Sigmoid function is used in the output layer.
- Input layer contains 77 neurons. For binary classification the output layer contains 1 neuron. The hidden layers contain 256, 128, 64 and 32 neurons respectively from first to fourth.
- The neuron count in each hidden layer decreases by half to ensure accurate output and reduce cost.

Cont..

- Between every two fully connected layers dropout of 0.1 and between output layer and last hidden layer dropout of 0.5 is used for the purpose of speeding up.
- Dropout removes neurons with their connection in a random manner to prevent overfitting and to speed up and reduce cost.
- For binary classification the Adam optimizer is used with default learning rate 0.001. Number of epochs is set as 20.
- The random forest classifier is used with 100 iterations as default value.

Dataset Description

- The dataset used here is CSE-CIC-IDS2018 dataset which is the result of a collaborative project between the Communications Security Establishment (CSE) the Canadian Institute for Cybersecurity (CIC).
- There are alomost 17 million data in the whole dataset, but the dataset I have used here has 331100 records after removing the null and infinity values.
- There are total 328181 normal packet flows and 92403 attacked packet flows. 1 is represented as normal and 0 as attacked packet in case of binary classification.
- There are 77 features that are used for classification.

Cont.

- There is a label for every record like Benign (i.e normal flow) or the name of the attack.
- The dataset includes seven different kind of attack like..
  - Bruteforce attack.
  - Infiltration attack.
  - DoS attack.
  - DDoS attack etc.
- The dataset is hosted in AWS and can be downloaded from there using the command line interface.

Preprocessing of Data

To supply more relevant data to the deep neural network classifier, a series of preprocessing operations are done on the data. These operations are as follows :

- Original dataset includes the source IP address,destination IP address,source port number and destination port number, which must be removed to ensure unbiased classification.Also timestamp information is removed as temporal information of the packet flow is redundant.

- Normalization is done on the dataset in each attribute between 0 to 1. This helps in providing more analogous values to the classifier.

- Every tuple with null or infinity value is dropped.

Experiment Details

- Due to the large size of the dataset the whole data cannot be loaded into a numpy array at once so I have done the experiment with the smallest file only.
- After the above mentioned preprocessing the data is split into train and test data into 80%and 20% respectively.
- After the training using Deep Neural Network and Random Forest,the performance of the classifiers are compared w.r.t different evaluation metrics.
- The confusion matrix and classification report are also computed using scikit-learn library.

Results and Evaluation

The performance of the Deep Neural Network classifier and Random Forest are evaluated using several evaluation metrics like accuracy, f1-score, recall and precision.
A confusion matrix is a representation of predicted results for a classification problem by providing the prediction of the samples separated as shown in Table-2.

	Class 1(Predicted)	Class 0(Predicted)
Class 1(Actual)	TP	FN
Class 0(Actual)	FP	TN

Table **2 :** Confusion Matrix

Cont

Then in terms of TP, FP, TN, FN the accuracy, recall and precision is defined as :

- Accuracy = (TP + TN)/(TP + FP + TN + FN) implies the fraction of correctly classified samples.
- Precision = (TP)/(TP + FP) implies the probability that a sample classified as positive is indeed positive.
- Recall = (TP + TN)/(TP + FN) implies the probability that the class is correctly recognized.

	Accuracy	Precission	Recall
DNN	77.95%	81.06%	77.94%
Random Forest	76.52%	75.09%	76.52%

Table **3 :** Binary Classification performance comparision

Concluion and Future work

This project aims to provide a model to simulate an IDS to correctly classify the normal and attacked packet flows.

- Comparison between two different models are done using evaluation metrics and we can see the DNN performs better in case of binary classification.

But there are several limitations of the model which can be overcomed in future.

- With the availability of large data and meaningful feature selection, I am hopeful of getting a very accurate result in near future with the whole data set.
- Multiclass classification was not possible with a fraction of dataset as the variety of attacks are not evenly distributed. So we can try multiclass classification techniques to accurately classify various attacks.

References

- Osama Faker and Erdogan Dogdu, “Intrusion Detection Using Big Data and Deep Learning Techniques,” In Proceedings of the 2019 ACM Southeast Conference (ACM SE ’19). Association for Computing Machinery, New York, NY, USA,2019, pp. 86–93.
- X. Zhang and J. Chen, ”Deep learning based intelligent intrusion detection,” 2017 IEEE 9th International Conference on Communication Software and Networks (ICCSN), Guangzhou, 2017, pp. 1133-1137.
- G. Karatas, O. Demir and O. Koray Sahingoz, ”Deep Learning in Intrusion Detection Systems,” 2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT), ANKARA, Turkey, 2018, pp. 113-116.
- G. C. Fernández and S. Xu, ”A Case Study on using Deep Learning for Network Intrusion Detection,” MILCOM 2019 - 2019 IEEE Military Communications Conference (MILCOM), Norfolk, VA, USA, 2019, pp. 1-6.


THANK
YOU