Anomaly detection optimization using big data.

Abstract:

Anomaly-based Intrusion Detection System (IDS) has been a hot research topic because of its ability to detect new threats rather than only memorized signatures threats of signature-based IDS. Especially after the availability of advanced technologies that increase the number of hacking tools and increase the risk impact of an attack. The problem of any anomaly-based model is its high false-positive rate. The high false-positive rate is the reason why anomaly IDS is not commonly applied in practice. Because anomaly-based models classify an unseen pattern as a threat where it may be normal but not included in the training dataset. This type of problem is called overftting where the model is not able to generalize. Optimizing Anomaly-based models by having a big training dataset that includes all possible normal cases may be an optimal solution but could not be applied in practice. Although we can increase the number of training samples to include much more normal cases, still we need a model that has more ability to generalize. In this research paper, we propose applying deep model instead of traditional models because it has more ability to generalize. Thus, we will obtain less false-positive by using big data and deep model. We made a comparison between machine learning and deep learning algorithms in the optimization of anomaly-based IDS by decreasing the false-positive rate. We did an experiment on the NSL-KDD benchmark and compared our results with one of the best used classifers in traditional learning in IDS optimization. The experiment shows 10% lower false-positive by using deep learning instead of traditional learning

Introduction:

Signature-based Intrusion detection systems are not suitable anymore to be used in nowadays network environment. Because signature-based models are not able to detect new threats and unknown attacks. Due to technology improvement, the number of attacks is increasing exponentially. Statistics show that attacks number increases with a rate of 100% each year causing huge money loss, about tens of millions of dollars for ransomware attacks only. Tis high number of millions of new threats that are developed every day, reduces the efectiveness of signature-based IDS because it is not a practical solution to update the signatures databases every few minutes. Anomaly-based IDS can be a better alternative of signature-based IDS because it is more suitable for nowadays challenges of new threats. Because it can detect new threats but still not practically used because of its high false-positive rate. Te reason behind the high false-positive rate is that anomaly model classifes an unseen pattern that did not learn it in normal cases, as an abnormal case. Te reason behind high false-positive is that the model has overftting what means it cannot generalize. Te solution may seem easy, by extending the training dataset to include all normal cases. But that still not a practical solution for a long time. Although we can add much more normal cases to datasets, we still need a model with higher ability of generalization. In this paper, we propose using deep learning with big data to solve this problem. Big data allows us to use big datasets for training to reduce the false-positive rate by including much more normal cases. And deep learning needs large datasets for training without facing the overftting problem as it may have more ability to generalize than traditional

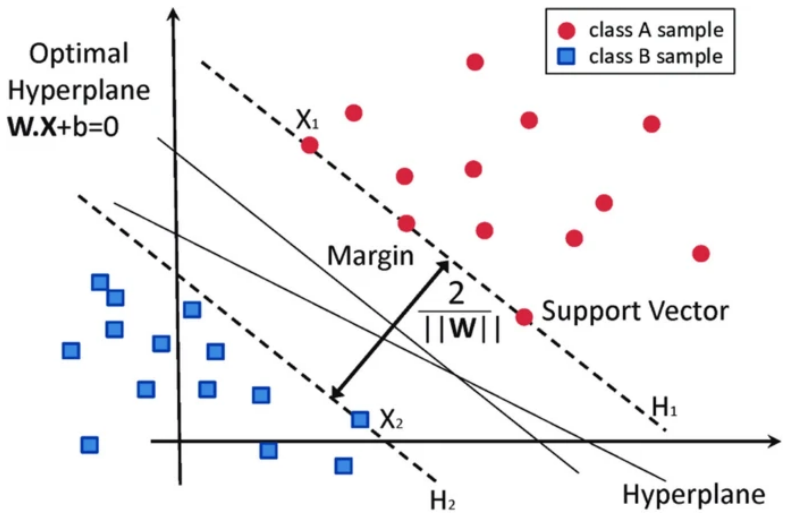
learning models. We compare results with one of the best traditional learning used classifers on the NSL-KDD benchmark dataset. Results show reducing false-positive to lower 10% than already used classifers.

Related work:

To optimize IDS, there are two general resolving directions. First way is to collect more data. Network security data and intrusion data are hard to collect because of data privacy concerns. In addition to the challenges of labeling data, which may be time-consuming for experts to do explanations and labeling process. Second way is studying how to increase the performance on small datasets, which is very important because the insights we can get from such researches can be implemented in big data researches. Tis paper is considered under the second way. We propose implementing deep learning instead of traditional learning. We choose to compare our works with SVM, shown in Fig. 1. Because SVM is the most popular traditional learning model used in network security and intrusion detection systems along years even in the era of big data. SVM has received a lot of interest in IDS optimization domain because it has proved by lots of experiments that it is one of the best classifers in anomaly-based intrusion detection systems

Method:

Te problem of any anomaly model is high false-positive rate, because it classifes any unseen pattern as abnormal where it may be a normal pattern but not included in the training set. Having a big training dataset that includes all possible normal cases has been considered as a theoretical idea that could not be applied in practice. Although we can increase the number of training samples to include much more normal cases, still we need a model that has more ability to generalize. We propose deep learning model instead of traditional learning as it may have more ability to generalize. Te proposed solution is to use a big data training set, but also to use a deep learning model to solve overfttting problem that occurs on the traditional learning models. We have done many experiments [26] that we care about collective and contextual attacks. Where we chose using a deep network with type of recurrent so we can handle sequences of actions or events. We chose LSTM to deal with sequences of inputs and focus on context. We were curious if LSTM can also achieve better results on small benchmark datasets or not. So we applied this experiment, we do not use sequences or even big data in this experiment. Only small benchmark data is used with LSTM.

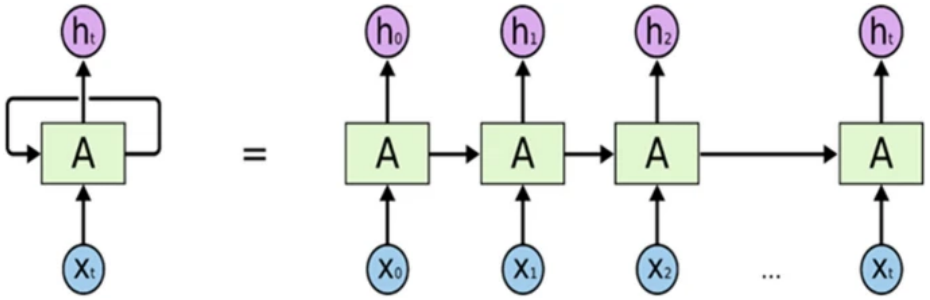


Proposed algorithm-long short term memory(LSTM):

LSTM is type of Deep Learning Recurrent Neural Network (RNN). see Fig. 2. RNN is extension of a convention feed-forward neural network. Unlike feedforward neural networks, RNN have cyclic connections making them powerful for modeling sequences. As a human no one think of each event separately. For example, when you are reading this article you read each word but you understand it in the context of this article, so that you understand the whole concept of this paper. Tat is the idea of RNN that has a loop to deal with input as a sequence, and that what we need to handle each event on network within its context.

Proposed algorithm-long short term memory(LSTM):

LSTM is type of Deep Learning Recurrent Neural Network (RNN).. RNN is extension of a convention feed-forward neural network. Unlike feedforward neural networks, RNN have cyclic connections making them powerful for modeling sequences. As a human no one think of each event separately. For example, when you are reading this article you read each word but you understand it in the context of this article, so that you understand the whole concept of this paper. Tat is the idea of RNN that has a loop to deal with input as a sequence, and that what we need to handle each event on network within its context



1. Long term dependency problem in RNNs.
2. Vanishing Gradient & Exploding Gradient

LSTM is designed to overcome vanishing gradient descent because it avoids long-term dependency problem. To remember information for long periods of time, each common hidden node is replaced by LSTM cell. Each LSTM cell consists of three main gates such as input gate it, forget gate ft, and output gate ot. Besides ct is cell state at time t. where xt, ht, and ct correspond to input layer, hidden layer, and cell state at time t. Besides, bi, bf, bc, and bo are bias at input gate, forget gate, cell state, and output gate, respectively. Furthermore, σ is sigmoid function. Finally, W is denoted by weight matrix

Experiment setup

T is experiment is performed on Google CoLab [30] using Keras libary (Python Deep Learning library) [31]. We apply LSTM of 64 hidden nodes with Relu activation function and

dropout=0.5. Using binary cross-entropy loss function. Using RMSprop optimizer. Learning rate=0.001, rho=0.9, decay=0.0.

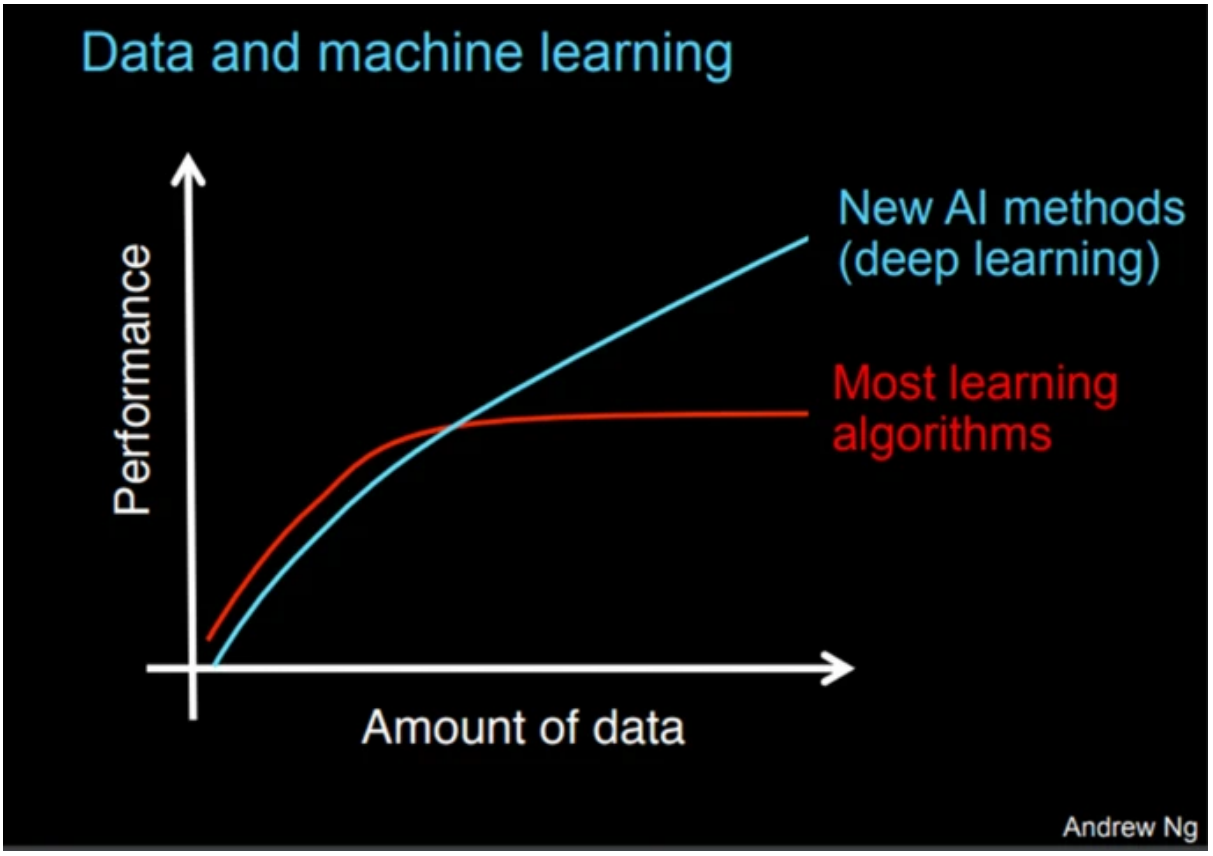
Data

KDD99

T is dataset is an updated version of the DARPA98 by processing the tcpdump portion. it was constructed in 1999 by the international competition, International Knowledge Discovery and Data Mining Tools Competition. Its size is 708 MB and it contains about 5 million connections [32]. It contains diferent attacks such as Neptune-DoS, pod-DoS, SmurfDoS, and bufer-overfow. Te benign and attack trafc are merged together in a simulated environment but it contains a large number of redundant records [33] KDD99 is the most famous research dataset but it cannot be used for real-life applications as data is old, not real, not enough.

Results and discussion

We will not discuss that big data is better with deep learning than traditional learning. It is a fact that the most important diference between deep learning and traditional machine learning is its performance as the scale of data increases. As we can see in Fig. 9 [36]. We did not use a big dataset in this experiment. We used a small dataset as we want to compare the results of applying deep learning instead of traditional learning in anomaly-based IDS even on small dataset. Tus, we can compare optimization of generalization. We got by experiment, that accuracy increases by using deep learning



model instead of traditional learning model. Moreover, false positive is getting lower by 10% less than traditional learning models. Results are shown in Table 5. We used SVM to compare with, because it has one of the highest results among traditional learning classifers. Te result we get is higher than all previous studies we have shown in related works.

Conclusion and future work:

Using big data analysis with deep learning in anomaly detection shows excellent combination that may be optimal solution as deep learning needs millions of samples in dataset and that what big data handle and what we need to construct big model of normal behavior that reduce false-positive rate to be better than small traditional anomaly models. We recommend using deep learning models instead of SVM when trying to build hybrid classifers for IDS. Because, as we see in results, deep models have more ability to generalize than traditional models. Tus, deep models achieve better results on unseen data causing less false-positive rate. We recommend using LSTM as we think it is an optimal solution in security domain because it can deal with sequences of events and context. So that it cannot only achieve better accuracy, but also can detect more various types of attacks that were not possible before. Such as, contextual and collective attacks.