2019 5th International Conference for Convergence in Technology (I2CT)
Pune, India. Mar 29-31, 2019

Anomaly Detection using Machine Learning Techniques

Sonali B. Wankhede
Assistant Professor, Institute of Computer Science
MET
Mumbai, India
sonaliw270@gmail.com

***Abstract*— Machine learning algorithms enable the systems to observe the behaviour based on real data. Based on the past experiences algorithms can be designed which allow computers to display behavior learned from past experiences. Machine learning algorithms are used to analyze the abnormal instances in a particular network. The algorithms can be trained for multiple data and can track the exploitation of a network. This idea is used for fraud detection and monitoring of the machines. Supervised learning is essential in terms of training and analyzing the abnormal behavior in a network. This paper presents the supervised techniques used to detect the network anomalies.**

Keywords— Supervised Learning, Anamoly Detection, MLP

I. INTRODUCTION

Anomaly detection is useful in identifying the patterns that are unexpected called outliers. It identifies the strange patterns in network traffic that would be a hack to system health monitoring (spotting a malignant tumor in an MRI scan). The machine learning algorithms are used to analyze the past attacks and develop appropriate defensive responses [1].

The algorithms analyze the directory traversals of a particular source or website and ensure the security standards of the web application services. Malicious websites directed through the destination path are also detected [1,2].

The algorithms for detection of path traversals can be used to identify various malicious domains. These algorithms can also detect the abnormal patterns from a host. Neural networks are useful for detecting unknown malicious attacks. The data sets are trained properly to analyze the behavior of the malicious or ransom files. A set of large number of ransom files and clean files are required for the training process.

The algorithms are used to identify the key features of every file in the data set. The features are further categorized into subsets to train the model for the acquired data set. When a system is attacked by an infected file, that file is checked by the trained model and actions are taken before the whole file system gets encrypted or locks the access to the computer.

A remote attack also known as remote exploitation is an attack where several network of computers are attacked. The system is accessed by the attacker through vulnerable points. The target of such type of an attack is exploitation and stealing sensitive data from the system.

Remote attacks are possible in the following ways:

- Denial of Service attack: The servers are flooded with large amount of false client requests in order to make the server unavailable to the user. The server is flooded with huge amount of pending client requests [14].
- DNS poisoning: The domain names like facebook.com are translated into some numeric IP addresses. In DNS poisoning the servers are basically tricked to accept the false data origins as legitimate and the users accessing the poisoned DNS servers are redirected to infected sites that downloads viruses to the system.
- Port scanning: The open ports on a network are identified by using a port scanner. The data is sent and received through computer ports. Hence the potential vulnerabilities are also detected. The attacker exploits the vulnerability by gaining access to the network.

The abnormal instances not correlating with the typical network behavior in a system can be identified and analyzed by the machine learning algorithms. Training can be done for multiple datasets to track the exploitation payload. Automation using algorithms, to learn from data and make predictions is allowed in machine learning. The algorithms can be supervised or unsupervised. In unsupervised learning the system automatically learns based on the updated data.

II. ANOMALY DETECTION

It is a mechanism for identifying the abnormal patterns in a network called as outliers.

It is useful in identifying the network traffic of strange pattern. Anomalies are broadly classified as:
Point Anomalies: The value of the data point is far outside the entirety of the data set in which it is found.

Contextual Anomalies: It is common in time series data. The abnormal behavior is context specific.

Collective anomalies: The instances of data are used in detecting anomalies.
The steps to build the algorithm are

1. The features x of the anomalous data are extracted
2. The parameters μ and σ are calculated
3. The probability p of x is calculated
4. Set probability boundary ϵ is tested

978-1-5386-8075-9/19/$31.00 ©2019 IEEE

By using a Gaussian distribution algorithm the sample x is distributed as follows:

$$\mu_j = \frac{1}{m}\sum_{i=1}^{m} x_j^{(i)}$$

$$\sigma_j^2 = \frac{1}{m}\sum_{i=1}^{m}(x_j^{(i)} - \mu_j)^2$$

The probability is calculated as follows:

$$p(x) = \prod_{j=1}^{n} p(x_j; \mu_j, \sigma_j^2) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(x_j - \mu_j)^2}{2\sigma_j^2}\right)$$

A real number evaluation metric is introduced after the implementation of the algorithm.
The steps to build the system are as follows:

1. Fitting the model p(x) on the training set
2. Predicting y on the resulting probabilities of the cross-validation and testing sets
3. Evaluating the result using a contingency table, precision/recall methods or the F1-score
4. Changing the values of ϵ (if required)

An anomaly detection system is used if the following conditions are satisfied:

- Availability of large number of negative examples
- Classification of anomalies is difficult and may vary in future datasets. e.g. fraud detection, monitoring machines, etc.

If there is large number of positive and negative samples with a similar range then it is beneficial to use a supervised learning algorithm. To analyze the errors the features are plotted and checked if the behavior is Gaussian. If it is not Gaussian then constants like log(x) can be added, to try to make it look as Gaussian as possible.

The basic assumption for using anomaly detection system is that, there is a need to have some of the anomalous as well as normal sample data.

III. SUPERVISED MACHINE LEARNING FOR ANOMALY DETECTION

A training set that contains both normal and anomalous sample data is required for constructing the predictive model. Supervised methods give better detection rate as compared to unsupervised methods.

The supervised techniques have capability of encoding interdependencies between variables, predicting events and ability to incorporate both prior knowledge and data [3].

A. Feedforward Neural Network

Feedforward neural networks are also called as Deep feedforward networks or Multilayer perceptron. In this network the function f* has to be approximated. A feedforward network defines the mapping and learns the values for parameters, which it gives best function approximation. The mapping is defined as y=f(x; θ) by learning the parameters.

In this network the information is transmitted from x through the computations to define f and to the output y. The outputs of the models are not fed back, as there are no feedback connections.

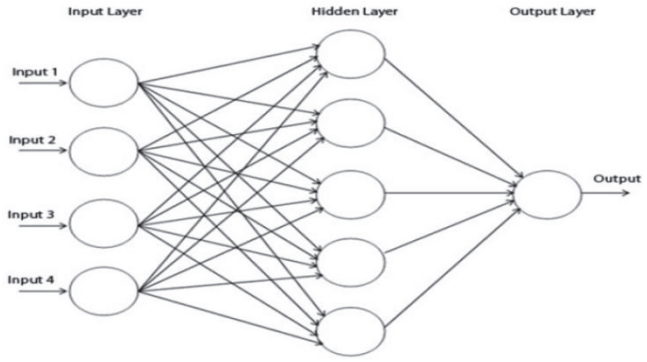

Fig. 1 Feedforward neural network

In perceptron learning if the weighted sum $\sum w_i x_i$ is less than the threshold value then the neuron output value is 0. If it is greater, then the output value is 1.The probability is predicted as an output. The sigmoid function exists between (0 to 1). It is used for the models where we have to predict the probability as an output. Therefore we can find the slope of the sigmoid curve at any given two points

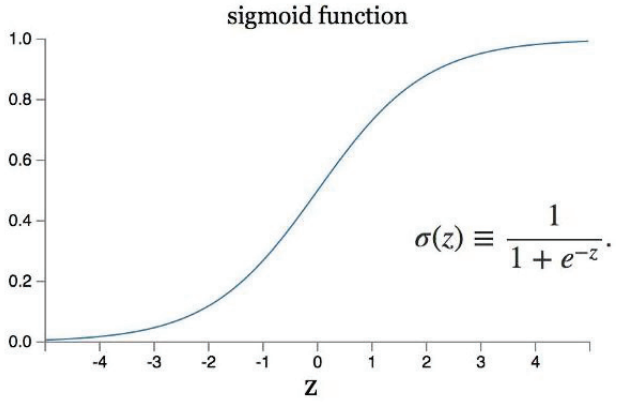

Fig. 2 slope of the sigmoid curve

Algorithm

Backpropagation algorithm is used to train the network. In order to reduce the error function a gradient descent method is used [4].

The input data is transmitted through the network until it reaches the output layer. Then the desired and actual outputs are compared and the error is calculated. The error is then back propagated. It gives the error for each neuron in the hidden layers. Using these values the algorithm updates the weights and biases. The goal of the algorithm is to make the output o_i as close to the desired output t_i for each input.

The error function is given by:

E=The error in the output layer k is calculated as:

$\Delta_k = t_k - o_k$

$\delta_k = \Delta_k \alpha'_k$

where α' is a derivative of the activation function. The weights in the output layer are given by:

$\Delta\omega_{jk} = x_k \delta_k \gamma$

Where x_k is the input from a neuron in the previous layer, γ is the learning rate. The learning rate regulates the speed at which the weights are adjusted.

IV. CONCLUSION

This paper presents the various types of Anomalies. Some of the remote attacks like DoS attack, DNS poisoning and Port Scanning are also discussed. Anomalies can be detected using the supervised machine learning technique like feedforward neural network.

RFERENCES

[1] S. B. Kotsiantis, P. E. Pintelas, I. D. Zaharakis, "Machine learning: a review of classification and combining techniques", Artificial Intelligence Rev, vol. 26, pp. 159-190, 2006.

[2] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification", Informatica (slovenia), vol. 31, pp. 249-268, 2007.

[3] Z. H. Zhou, "Rule extraction: Using neural networks or for neural networks?", Journal of Computer *Science and Technology*, vol. 19, no. 2, pp. 249-253, 2004.

[4] Demuth, Howard B., et al. Neural network design. Martin Hagan, 2014.

[5] Sang-Jun Han, Sung-Bae Cho, "Evolutionary Neural Networks for Anomaly Detection Based on the Behavior of a Program", IEEE Transactions on Systems, vol. 36, no. 3, June 2006.

[6] V Theuns, H Ray, "Intrusion Detection and Approahes", Journal of Computer Communications, vol. 25, pp. 1356-1365, 2002.

[7] A Alfantookh, "DoS Attacks Intelligent Detection using Neural Networks", Journal of King Saud University Computer and Information Sciences, vol. 18, 2005.

[8] J. F. Nieves, Y. C. Jiao, "Data clustering for anomaly detection in network intrusion detection", Research Alliance in Math and Science, pp. 112, 2009.

[9] Yusuf Sani. "An overview of neural networks use in anomaly Intrusion Detection Systems", IEEE Student Conference on Research and Development (SCOReD), 2009.

[10] Zhu Xiaojin, Semi-supervised learning literature survey, 2005.

[11] A. R. Barron, "Universal Approximation Bounds for Superposition of a Sigmoidal Function," IEEE Trans. Information Theory, vol. 39, pp. 930-945, 1993.

[12] T. Kwok, and D. Yeung, "Constructive Feedforward Neural Networks for Regression Problems: A Survey," Technical Report of the Computer Science Department, Hong Kong University of Science and Technology, HKUST-CS95-43, 1995.

[13] D. A. Ludeña R., S. Kubota, K. Sugitani, Y. Musashi: DNS-based Spam Bots Detection in a University, International Journal of Intelligent Engineering and Systems, Vol. 2, No. 3, 2009, pp.11-18.

[14] Sonali B. Wankhede, "Study of Network-Based DoS Attacks", Proceeding of Nanoelectronics, Circuits and Communication Systems, pp 611-616, 2017.

[15] J. Mirkovic, P. Reiher, "A taxonomy of DDoS attack and DDoS defense mechanisms", ACM SIGCOMM Computer Communication Review, vol. 34, no. 2, pp. 39, 2004.

[16] I. Witten, E. H Frank, Practical Machine Learning Tools and Techniques Second Edition, USA:Morgan Kaufmann Publications, 2005.

[17] M.F. Amasyah, New Machine Learning Methods and Drug Design Applications, 2008.