International Journal of Engineering & Technology, 7 (2.20) (2018) 86-89

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Secure automated threat detection and prevention (SATDP)

CH. Ramaiah *, D. Adithya Charan, R. Syam Akhil, P. Pavan Kumar

[1]*Computer Science, Koneru Lakshmaiah Education Foundation, Guntur, India.*
**Corresponding author E-mail:ramaiah.challa@kluniversity.in*

Abstract

Secure automated threat detection and prevention is the more effective procedure to reduce the workload of analyst by scanning the network, server functions& then informs the analyst if any suspicious activity is detected in the network. It monitors the system continuously and responds according to the threat environment. This response action varies from phase to phase. Here suspicious activities are detected by the help of an artificial intelligence which acts as a virtual analyst concurrently with network intrusion detection system to defend from the threat environment and taking appropriate measures with the permission of the analyst. In its final phase where packet analysis is carried out to surf for attack vectors and then categorize supervised and unsupervised data. Where the unsupervised data will be decoded or converted to supervised data with help of analyst feedback and then auto-update the algorithm (virtual analyst). So that it evolves the algorithm (with active learning mechanism) itself by time and become more efficient, strong. So, it can able to defend form similar or same kind of attacks.

Keywords*: Artificial intelligence, intrusion detection System, network security, machine learning, (supervised and unsupervised) learning.*

1. Introduction

With the recent advancements in network technology and growth of internet usage so the attack vectors. Accord to stats Q3 2016 alone 18million new malware samples were captured. we are going to reduce the threats by implementing AI (artificial intelligence) in cyber security field where we can prevent most of the attacks. Through Signature and behavior basis we can reduce attack on network and server fields by implementing AI. This all can be done on basis of deep learning and machine learning. In this we train a machine how to detect and prevent the attacks in the server network or the system which is a supervised learning. We are introducing Remote Access tab concept which is used by analyst and he can operate the work from anywhere in emergency times as the analyst is the only admin. And by the attacks which are predefined are automatically removed by that we reduce the burden. If there are new attacks or any malicious packets being observed entering the server in the firewall the virtual analyst tries to detect it or by unsupervised data, it try's removes the attack or prevents it from entering the network and notify the analyst.

1.1. Artificial intelligence

Is an area of computer science that emphasizes the creation of intelligent machines that works & react like humans(by implementing different type of feasible techniques uses to solving problems, planning, learning, managing, recognition) which involves deep learning, machine learning.

1.2. Deep learning concepts of AI

Deep learning or also known as active learning which enables the AI to analyze the current scenario which isn't possible in any other types of intrusion detection system. So with the evolution modern technology in IT has given rise to the birth of more complex systems and networks[15] as well as increased threats.[5][4] we can overcome these by implementing an AI in networks through its deep learning capability which is quite popular now-a-days has been successfully implemented in advanced analytic algorithms. Now it's time for networks we can use this AI as an virtual analyzer which analyzes the behavior to determine attack vectors that bypass IDS systems.

2. Machine learning

Machine learning falls under the domain of artificial intelligence. These are used to make predictions in the data. Machine learning algorithms should be trained in order to make predictions. Training means initially the algorithms has to be exposed to examples These examples can range from several thousands. The algorithm has to run with example datasets before handling real data sets. This training period is also known as learning phase that enables the Algorithm to increase its efficiency and reliability over real time scenarios. This learning has two categories supervised learning, unsupervised learning. In supervised learning Algorithm is trained using labeled data? Labeled data means the input and output will be present with corresponding relation. Unsupervised learning the algorithm uses a training set consists [14] of unlabeled data which means the algorithm should interpret the corresponding relation between input and output.

2.1. Supervised learning (signature)

It means a predefined dataset where we feed the data of previous attacks or malicious packets. Which have the data of different type of classifications on basis of behavioral analysis [9] (big data). Here we find all previous attacks if update by analyst. It acts more efficient. Where accuracy will be high and easily filtered out. Here

we have botnets, malicious packets, spoofed packets data in a dataset. Same function as signature based algorithm. Here we use decision tree classifier.

2.2. Semi-supervised learning (TSVM)

It produces dataset where (tsvm) is a transductive support vector machine (it is semi supervised method which uses of unlabeled data together with labeled data to build classification)

2.3. Unsupervised learning (anomaly based detection)

Here it uses unlabeled data[7] which means the packet [1] pattern or signature is not defined earlier. For detecting a undefined data or definition we uses KNN (k nearest neighbor & svm-support vector machine algorithm).

3. Machine learning in IDS

*(*using different types of algorithms) [14]The main problem is that they are evolving. Over ten thousand of new signatures are uploaded for analysis every day. Also, one of the draw-back of current IDS is their manual maintenance. Using Machine learning techniques, we can improvise this scenario. Instead of manually analyzing the signatures and updating them to IDS database using unsupervised learning technique we can train an AI for cyber defense purposes. Since intrusion detection is mostly based on classification. Certain issues may arise when the Machine Learning algorithms such as network diversity, detection of new attacks, false positives rate, learning. network diversity is one of the major issues to be concerned because normal network traffic is difficult to define and varies over bandwidths, behavior, etc. since the algorithm initially acts as an signature based IDS the algorithm may find a net threat close normal curve but this can be reduced as time progress the algorithm familiarizes with the environment. [11]One biggest problem is that to teach the algorithm about the malicious behavior as these tend to differ from one malicious anomaly to another without proper generalization there may be rise in false positive rate. The data has to be processed before it fed into the algorithm means the features have to be chosen carefully some can be found initially while others can be determined by diagnosis of algorithm.

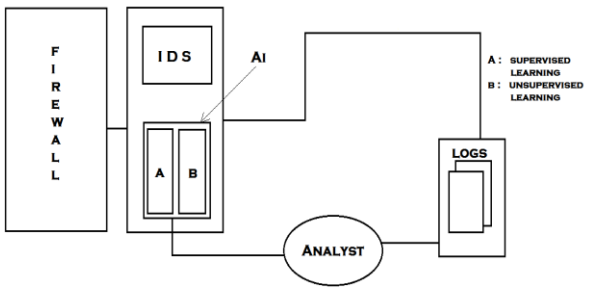

Fig. 3.1: Working of AI in IDS

4. Data used here

Table 4.1: Data Description or Data Types are Used Here

Id	Duration(hrs)	# Packets	#NetFlows	Size	Bot	#Bots
1	6.15	71,971,482	2,824,637	52GB	Neris	1
2	4.21	71,851,300	1,808,123	60GB	Neris	1
3	66.85	167,730,395	4,710,639	121GB	Rbot	1
4	4.21	62,089,135	1,121,077	53GB	Rbot	1
5	11.63	4,481,167	129,833	37.6GB	Virut	1
6	2.18	38,764,357	558,920	30GB	Menti	1
7	0.38	7,467,139	114,078	5.8GB	Sogou	1
8	19.5	155,207,799	2,954,231	123GB	Murlo	1
9	5.18	115,415,321	2,753,885	94GB	Neris	10
10	4.75	90,389,782	1,309,792	73GB	Rbot	10
11	0.26	6,337,202	107,252	5.2GB	Rbot	3
12	1.21	13,212,268	325,472	8.3GB	NSIS.ay	3
13	16.36	50,888,256	1,925,150	34GB	Virut	1

Table 4.2: Distribution of Labels in the NetFlows for Each Scenario in the Dataset

Scen.	Total Flows	Botnet Flows	Normal Flows	C&C Flows	Background Flows
1	2,824,636	39,933(1.41%)	30,387(1.07%)	1,026(0.03%)	2,753,290(97.47%)
2	1,808,122	18,839(1.04%)	9,120(0.5%)	2,102(0.11%)	1,778,061(98.33%)
3	4,710,638	26,759(0.56%)	116,887(2.48%)	63(0.001%)	4,566,929(96.94%)
4	1,121,076	1,719(0.15%)	25,268(2.25%)	49(0.004%)	1,094,040(97.58%)
5	129,832	695(0.53%)	4,679(3.6%)	206(1.15%)	124,252(95.7%)
6	558,919	4,431(0.79%)	7,494(1.34%)	199(0.03%)	546,795(97.83%)
7	114,077	37(0.03%)	1,677(1.47%)	26(0.02%)	112,337(98.47%)
8	2,954,230	5,052(0.17%)	72,822(2.46%)	1,074(2.4%)	2,875,282(97.32%)
9	2,753,884	179,880(6.5%)	43,340(1.57%)	5,099(0.18%)	2,525,565(91.7%)
10	1,309,791	106,315(8.11%)	15,847(1.2%)	37(0.002%)	1,187,592(90.67%)
11	107,251	8,161(7.6%)	2,718(2.53%)	3(0.002%)	96,369(89.85%)
12	325,471	2,143(0.65%)	7,628(2.34%)	25(0.007%)	315,675(96.99%)
13	1,925,149	38,791(2.01%)	31,939(1.65%)	1,202(0.06%)	1,853,217(96.26%)

Data sets can be categorized into 3 types when it comes to machine learning. They are labeled data, unlabeled data[8][7], and semi labelled data. Labeled data as the name suggests the data samples were assigned or marked with a label or class to which they belong. The samples may have more than one label attributes. These are fed into supervised machine learning algorithms for training. Unlabeled data sets are raw datasets in this the data samples doesn't contain any labels or any other to differentiate among each other. Consider it as a dataset which is unclassified or unsorted data. These are generally classified using unsupervised learning algorithms. Semi labeled dataset is special kind of dataset as it contains both labeled dataset and unlabeled dataset. Labeling a large dataset can be a tedious task to overcome this a part of dataset is taken and it is labeled using a model or another labelled data set of same category is taken and now the unlabeled data is pseudo labelled by the model with the help of labeled data. Again now both of the pseudo and labeled data sets are fed to retrain the model and labeled data is achieved as result.

5. Algorithms Used In This Project

5.1. KNN

K Nearest Neighbor classifier is an supervised ML algorithm mostly used for regression and classification predictive problems. The main reason to choose these algorithm is

- Calculation time
- Predictive power
- Easy to interpret output

In this project we are using it for the classification of the training data from its classes. For a better understanding let us assume a data set and project it on a 2-dimensional space now in order to classify the value of k is taken and it no of neighbors near it and it belongs to class which has more elements present near to it as shown in the figure. For computation of distance between data

various methods are Used. For a continuous data Euclidean distance is used as for discrete variables another metric hamming distance and other methods are used. For a continuous data Euclidean distance which computes distance using voronoi cells are used as for discrete variables another metric hamming distance and other methods are used.

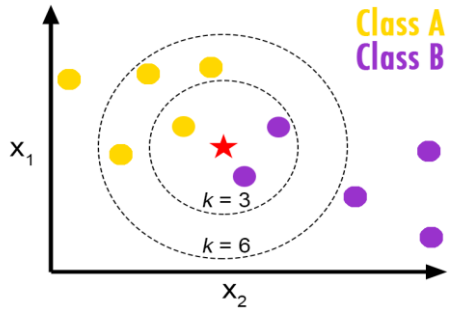

Fig. 5.1: KNN

For deeper understanding let us see the above figure for different values of k the unclassified data can be classified into classes which changes as k changes. However skewed data can be a problem for KNN. It is type of data present in data set that which belong to an underrepresented class in the complete dataset. As the classification is done on the basis of most nearest class popular or dense classes may mislead the prediction. To overcome this weights can be computed proportional to inverse of distance which I similar to 1/distance(a,b) where a is the data point for which is being classified and b is its neighbor. In the end the choice of k matters. The higher values of k reduces noise while predicting. However larger values of k results in less distinct boundaries between classes. Irrelevant features in dataset can also reduce the accuracy of the algorithm. Evolutionary models can be used to overcome these.

5.2. TVSM and SVM

Mostly support vector machines are used for classification and regression analysis. One class SVM[13] (support vector machines) is used handling unlabeled data as it uses unsupervised learning technique. Basically SVM's represent data from data sets as a points in space and each category are separated by vectors or a plane. It acts a boundary which is also referred as decision boundary between classes. Data labeling is done based on relative position of data in the space.

The prominent features of SVM is it uses a non-linear function to project the data creating the non-linear decision boundary. Which means the data which can't be labeled in original space are lifted to feature space where a hyperplane separates or labels them.

According to Tax and Duin approach the model takes a spherical boundary instead of planar in feature space. This hyper sphere is shrunk to minimize the effect of absorbing outliers in the solution. The hypersphere contains center a linear combination of all support vectors and radius r as distance between center to boundary of which a volume of R^2 is minimized. The distances from data point's x_i is less than r and create a soft boundary or margin with slack variables ξi with penalty parameter C. Then the minimization problem becomes After computation of above a new data point can classified in or out of class using Gaussian-Kernel distance function between 2

$$\min_{R,\,\mathbf{a}} R^2 + C\sum_{i=1}^{n}\xi_i$$

subject to:

$$\|x_i - \mathbf{a}\|^2 \le R^2 + \xi_i \qquad \text{for all } i = 1,\ldots,n$$
$$\xi_i \ge 0 \qquad \text{for all } i = 1,\ldots,n$$

Fig. 5.2: KNN

5.3. Decision tree

Decision tree classifier is a supervised machine learning algorithm. As the name suggests this algorithm classifies the data by breaking it down into a tree until leaf nodes are reached. It generates the tree by choosing an attribute or feature among data [3]and splitting according to them which results in in pure and impure sets. Where the pure set is the leaf. for each and every level lt computes the entropy to form the child nodes fromparent. It revolves around what feature or attribute to choose and when to choose.

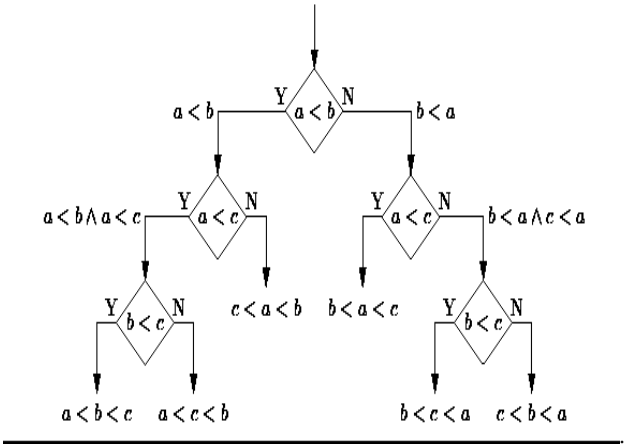

Fig. 5.3: Decision tree functioning

6. Implementation

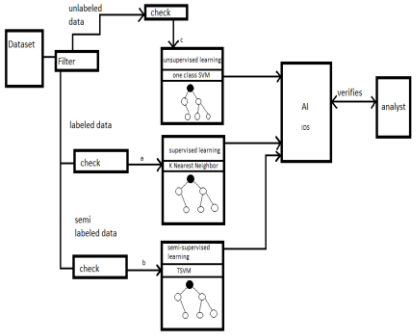

Fig. 6.1: Process flow diagram

a. Here data stored will be given or forwarded for analysis where data means malicious, normal and damaged data.
b. Dataset: filter
c. Where filter will compare the data with predefined precision or definitions then classified the data and forward to respective algorithms case as mentioned in diagram.
d. Data we provided to system (artificial intelligence), here data is used to train the AI(it is a combination of machine learning and deep learning in it by using different algorithms.)
e. From here data will forwarded to algorithm for analysis and training. As mentioned in earlier according to data set type we choose the algorithms and place it in main program.
f. It use KNN (K-Nearest neighbor) for supervised data set for quick and best analysis.
g. The steps and functions involved in it is explained at 5(a) section. For it quick predictive and classifies. We use output values compared to targeted values easily and re-assign the algorithm for best expected output. By that it builds a hierarchical mapping so can we get quick result next time.
h. It uses the TSVM algorithm used here for semi labeled data. Where it compare with labeled data with unlabeled data then predict the values and train the data mapping and convert the unlabeled to labeled data and forward to analyst for more assistance.

i. It uses the SVM or ONE CLASS SVM one class support vector machines algorithm for unlabeled data. it uses unsupervised learning technique. Basically SVM's represent data from data sets as a points in space and each category are separated by vectors or a plane. It acts a boundary which is also referred as decision boundary between classes. Data labeling is done based on relative position of data in the space. Then map the data and forward analyzed values to analyst for further analysis and verification.

j. here if any error occurs it start form beginning again same process will repeat.

k. As mentioned in above steps where output data will be categorized and mapped in analysis new labeled layer network by that the next time scan or analysis we will be quick then previous as the previous scan already build a basic layer

l. Then we compare data (output data) with the targeted values or defined value if it not precise then we re-assign and retrain the algorithms with same dataset until we get optimal or desired values.

m. By that re-assigning and re-training the algorithms with same dataset we will get the development in its performance by comparing values with desired or targeted values.

n. By using different dataset or data models we can expect more negative values to occur. So we can reduce it by repeating the functions or process until we get desired or required value.

7. Results

	KNN	Decision tree classifier	One class svm
True Positive	28379	29180	6987
True negative	1734177	1673382	50890
False positive	43884	104676	87738
False negative	1682	881	80852
recall	0.944047104221	0.970692924387	0.0795432552739
Precision	0.392718265226	0.217990572169	0.0737608867775
Fscore	0.554689027012	0.356027330405	0.0765430205298

Fig. 7.1: Result for Precision, Recall, fscore.

8. Conclusion

Artificial Intelligence (AI), Machine Learning and Big Data Behavioral analytics in cyber-security plays a critical role. The modes of learning i.e. supervised and unsupervised documents the threats and auto-update (active learning segment) the virtual analyst algorithm with permission of analyst to shield the network. By that it reduces the workload of the analyst and in less time, we get more efficient system, so we can minimize the attack impact on institution or an organization network. So, by that analyst can more focus on research or development side of network.

References

[1] Chandola V, Banerjee A & Kumar V, "Anomaly detection: A survey", *ACMComput. Surv.*, Vol.41, No.3, (2009), pp.15:1–15:58.

[2] Aggarwal CC, "Outlier ensembles: Position paper", *SIGKDD Explor. Newsl.*, Vol.14, No.2, (2013), pp.49–58.

[3] ling Shyu M, ching Chen S, Sarinnapakorn K & Chang L, "A lanomaly detection scheme based on principal component classifier", *Proceedings of the IEEE Foundations and New Directions of Data Mining Workshop, in conjunction with the Third IEEE I international Conference on Data Mining*, (2003), pp.172–179.

[4] Hawkins S, He H, Williams G & Baxter R, "Outlier detection using replicator neural networks", *Data Warehousing and Knowledge Discovery, ser. Lecture Notes in Computer Science*, Springer Berlin Heidelberg, Vol.2454, (2002), pp.170–180.

[5] Scholz M & Vigário R, "Nonlinear PCA: a new hierarchical approach", *Proceedings of the 10th European Symposium on Artificial Neural Networks (ESANN)*, (2002), pp.439–444.

[6] Schubert E, Wojdanowski R, Zimek A & Kriegel H, "On evaluation of outlier rankings and outlier scores", *Proceedings of the Twelfth SIAM International Conference on Data Mining*, (2012), pp.1047–1058.

[7] Zimek A, Campello RJ & Sander J, "Ensembles for unsupervised outlier detection: Challenges and research questions a position paper", *SIGKDD Explor. Newsl.*, Vol.15, No.1, (2014), pp.11–22.

[8] Pelleg D & Moore AW, "Active learning for anomaly and rare-category detection", *Advances in Neural Information Processing Systems*, (2004), pp.1073–1080.

[9] Yen TF, "Detecting stealthy malware using behavioral features in network traffic", *Ph.D. dissertation, Carnegie Mellon University*, (2011).

[10] When Big Data Met Security: Is The New Era Beginning? Chuck Hollis, VP – CTO, EMC Corporation, 2011.
http://chucksblog.emc.com/chucks_blog/2011/08/when-big-data-met-security-is-the-new-era-beginning.html

[11] Vijayarani S &Dhayanand S, "Liver Disease Prediction using SVM and Naïve Bayes Algorithms", *International Journal of Sci-ence, Engineering and Technology Research*, Vol.4, (2015), pp.816-820.

[12] Wang J, Jebara T & Chang SF, "Semi-supervised learning using greedy max- cut", *Journal of Machine Learning Research*, Vol.14, No.1, (2013), pp.771-800.

[13] Utkin V & Zhuk YA, "An one-class classification support vector machine model by interval-valued training data", *Knowledge-Based Systems*, Vol.120, (2017), pp.43-56.

[14] Chang A, "R for Machine Learning, Prediction: Machine Learning and Statistics", *MIT OpenCourseWare*, (2012), pp.1-8.

[15] Sharma V, Rai S & Dev A, "A Comprehensive Study of Artificial Neural Networks", *International Journal of Advanced Researchin Computer Science and Software Engineering*, Vol.2, No.10, (2012), pp.278-284.