IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-6, Issue-12, Dec 2018 E-ISSN: 2347-2693

Hybrid Distributed Intrusion Detection System

A.A. Ujeniya[1*], R.D. Pawar[2], S.A. Sonawane[3], S.B. Shingade[4], S.R. Khonde[5]

[1] Department of Computer Engineering, Modern Education Society's College of Engineering, Pune, India
[2] Department of Computer Engineering, Modern Education Society's College of Engineering, Pune, India
[3] Department of Computer Engineering, Modern Education Society's College of Engineering, Pune, India
[4] Department of Computer Engineering, Modern Education Society's College of Engineering, Pune, India
[5] Assistant Professor, Modern Education Society's College of Engineering, Pune, India

**Corresponding Author: adityauj@gmail.com, Tel.: +91-8888358021*

Available online at: www.ijcseonline.org

Accepted: 13/Dec/2018, Published: 31/Dec/2018

***Abstract*—** There is rise in new Intrusion Detection Systems (IDSs) due to increasing frequency of various malicious activities over network and certain network policy violations. IDS, being an advanced tool and equipment to secure the network parameter by surveillance from the different network risks, is capable of detecting various attacks due to advancements in Computer Science. These advancements include machine learning models which can be integrated into an IDS for increasing the Detection Rate of attacks and minimizing the False Alarm Rate (false positives). In this paper, Hybrid Distributed IDS (HDIDS) is proposed in which strengths of Signature-based and Anomaly-based detection are combined together to detect different types of Denial of Service (DoS) attacks. HDIDS is presented by combining an anomaly-based detection algorithm and multiple signature-based detection algorithms. The signature-based multiple classifiers ensemble and can detect real time attack based on majority of votes from each classifier. Ensembled output use voting technique which are simplest to implement and produce favourable results. Anomaly based classifier has intensive focus over new and unknown attacks in distributed network. The dataset used for training the classifiers is ISCX CICIDS-17 consisting of latest attacks and 88 features providing better options for feature selection with respect to each classifier.

***Keywords*—** Machine Learning, Hybrid, Intrusion Detection System, Anomaly-based classifier, Signature-based classifier, Ensemble

I. INTRODUCTION

The role of internet and computer networks in everyday activities of human lives lead to advancement in usage and development of internet related applications. It has a global impact and provides worldwide platform to various users and organizations to share and store confidential information. As a result, cyber security is currently a huge area of research, as it has a large impact over network community. Considerable security mechanisms such as firewall, user authentication, antivirus and access control are developed to protect individual host or computer network from various abnormal activities or potential malicious attacks performed by the intruder.

There is serious need to put something on line of defense against such attacks in spite of various security mechanism available. Protecting user system is also a main issue that is tremendously increasing nowadays. Intrusion Detection Systems (IDSs) is a software that automates the intrusion detection process and expected to minimize the influence of such potential attacks. IDS is used as security improvement to protect the user as well as network from getting compromised by attacks. Exploiting policies like Confidentiality, Integrity and Availability or even penetrating the security mechanism of computer system is handled by IDS.

Also, IDS can monitor network activity by probing packet data for detecting attacks. Based on scrutiny performed, there are 2 different approach in intrusion detection methods. Anomaly-based detection system are used to identify unknown attacks by deviating from normal activity. Whereas, Signature-based detection system is used to detect known attacks from predefined patterns of the attacks. A dataset is required to have knowledge about the attacks, while anomaly systems use only packet information to detect attack. Both methods have their own strengths and weakness. Anomaly-based systems may outperform signature-based systems in terms of detecting unknown attacks. It is not possible to define all attacks without knowing all the attack patterns beforehand. So, does it have certain disadvantage. It has more False Alarm Rate (FAR) that results in flagging a normal attack as malicious activity. Here, signature-based systems come in handy. Due to predefined patterns, the ability to distinguish attacks from normal behavior results in reduced FAR and improved Detection Rate (DR).

To complement each other's strengths, hybrid IDS has been developed where IDS contains both signature-based systems and anomaly-based systems. Hybrid approach would give top-notch performance based on method of combination.

In this paper, we propose Hybrid Distributed IDS in which 1 anomaly-based and 2 signature-based systems will be used. Signature-based systems will use Supervised Learning algorithms and anomaly-based system will use Unsupervised Learning algorithm. All the known attacks from the dataset will be trained onto proposed supervised algorithms – Decision Tree, k-Nearest Neighbor. New and unknown attacks detected by the unsupervised learning algorithm is captured and sent to main agent. The main agent will try to uncover whether the abnormal activity suspected by anomaly-based system is FAR or an attack. If it is an attack, a signature for the new attack will be created and added to dataset. All the supervised algorithms will be trained again for detecting the new attack in near future. Furthermore, the output of attacks detected by supervised algorithms will be ensembled together based on majority of voting. This will help to correctly classify attack and reduce FAR. The dataset used is ISCX CICIDS-2017 which consists of all the latest attacks. From the dataset, we are targeting only Denial of Service DoS attacks for building efficient Hybrid IDS.

Rest of the paper is organized as follows, Section I contains the introduction of Hybrid Distributed Intrusion Detection System, Section II contain the related work of various IDS build using signature-based, anomaly-based or hybrid of both. Section III contain the detailed description about the latest dataset used for training the classifiers, Section IV contain the components and essential information about the various classifiers used and Section V concludes research work with future directions).

II. RELATED WORK

Multi-Classifier systems have received much attention due to their focus on combining the output of classifiers and create a final decision. To accomplish best possible results, the hybrid IDS's come into action by combining signature-based systems and anomaly-based systems.

Chun Guo et. Al. proposes a 2-level hybrid IDS. In this, they have 2 anomaly-based detecting component and 1 signature-based detecting component. They have implemented k-Nearest Neighbour to develop signature-based system and K-means clustering for anomaly-based system. The model proposed by them has drastic reduction in FAR, down to 1.05% even for new attacks. Their model's accuracy is 95.76% which they have tested on KUBA dataset [1].

Georgios P. Spathoulas and Sokratis K. Katsikas has performed the reduction of False Alarm Rate I.e. False Positives. Most of the anomaly-based systems have higher FAR. Using this approach, it is possible to reduce FAR. They have 3-level components – Neighbouring Related Alerts, High Alert Frequency and Usual False Positives. Combing these 3 components, they check SourceIP and calculate the time interval for which the SourceIP is sending data. Also based on size of packet and time interval, their filter is able to distinguish normal behaviour with malicious attacks. The filter has ability to reduce FAR by 90% as mentioned in the results [2].

Pedro Casas, Johan Mazel, Philippe Owezarski has proposed an anomaly-based IDS which does not use any dataset to train their algorithm. It is constructed in 3 steps. First is Multi Resolution Flow Aggregation method in which all the anomalous marked packets are flagged for further inspection. In second step, the packets marked anomalous are ranked according to their abnormality by using Sub Space Clustering method. In third step, the top ranked outlying flows are flagged anomalous using simple threshold detection approach. If the outlier lies beyond threshold specified, it is declared as attack. This proposed method has reduced FAR and increased accuracy, both benefiting at the same time. Their model was observed to have higher accuracy as compared to signature-based learning algorithms [3].

Hadi Sarvari and Mohammad Mehdi Keikha paper proposes combination of machine learning approaches is used to detect the system attack. This paper also provided solution to unbalanced data for some classes of data. The best combinatory model is one containing SVM, DT, 1NN, 2NN and 3NN. Output of all the models were feed in neural network for further classification which will uplift the Detection Rate. Their accuracy for Normal behaviour - 99.2%, DOS attack - 98.21%, Probe attack - 93.22%, U2R attack - 44.44% and R2L attack - 93.21% [4].

Alex Shenfield, David Day and Aladdin Ayesh propose a mode in which artificial neural networks are used to detect malicious network traffic suitable for deep packet inspection. Proposed network has 1000 input nodes, 30 nodes in 1st Hidden layer 1, 30 nodes in 2nd Hidden layer nodes and 2 output nodes. The proposed artificial neural network architecture is able to distinguish between normal and malicious network traffic. Results for their model were: Accuracy: 98%, Precision: 97%, Sensitivity: 95% [5].

Preeti Aggarwal and Sudhir Kumar Sharma cleared our perspective of how to look at the attributes, they taken NSL-KDD for research purpose. WEKA tool with Random forest is used for the simulation of the dataset. Preeti Aggarwal *et al* chose tree-based algorithm for the simulation as the provide better accuracy and precision than other algorithms. The dataset NSL-KDD is clustered into 4 classes Basic, Content, Traffic, Host. This research helped us to choose the dataset, and perform the pre-processing on out dataset (CICIDS 2017) [6].

Rana Amir Raza Ashfaq *et al* proposed a model for training the IDS which includes their own algorithm. According to their research the DR and FAR both increased and decreased respectively. A single hidden layer feed-forward neural network (SLFN) is used to train to the output which is classified as low, mid and high depending upon the result. The algorithm is applied to KDD dataset and split into (10,90) 10% for training and 90% for testing as KDDTest^{+} and KDDTest^{-21}. SLFN given an accuracy of 84.12 and 64.82 in KDDTest^{+} and KDDTest^{-21} respectively. They investigated divide and conquer strategy in which unlabelled samples with their predicted labels are categorized according to decided magnitude [7].

Sandhya Peddabachigari et Al. has proposed a model for IDS using intelligent systems where Decision Tree (DT) and Support Vector Machine (SVM) combine as hybrid classifier system. Ensemble learning is performed for combining the output of both classifiers. We learned that their classifiers have the ability to detect and quickly respond to the anomalous behaviour. But it has its inability to detect some intrusions if certain particular sequence of event has not been recognized and furthermore not creates specific rules for it. They evaluated their model on KDD Dataset. Their model has 100% accuracy rate for Probe class whereas the accuracy for other classes can be lifted if proper base classifiers can be used [8].

S. Khonde and V. Ulagamuthalvi proposed an ensembled system where they have used various supervised and unsupervised algorithms where the output of every algorithms is ensembled to classify the attack accordingly. They have used NSL-KDD dataset for training and proposed to work in Realtime Environment [9].

P. Rutravigneshwaran proposed a study of IDS that uses KDD 99 Winner Cup Dataset for evaluating the latest attacks like U2R and R2L. He also mentions 2 algorithms that can perform better than rest of the supervised algorithms - Decision Tree (DT) and Support Vector Machine (SVM). In his study, it is found that DT overall performs faster and better than SVM and Fast Hierarchical Relevance Vector Machine (FHRVM) [10].

III. DATASET – ISCX CICIDS - 2017

By observing the existing dataset since 1998, they are found out to have less volume and less up to date attacks. As the today's world is facing some problems with increasing exchange in data sharing within network, gives a rise to different potential attack. To design a sophisticated IDS knowing all new attacks with different functionality there is a need of fully adequate dataset. There exist some dataset including DARPA 98, KDD 98, ISC 12, ADFA 13, do not trend the new attack techniques used today, which can be termed as outdated and unreliable. CICIDS contains benign and seven common attacks which reflects the current trends. CICIDS2017, which covers all the eleven criteria with common updated attacks such as DoS, DDoS, Brute Force, XSS, SQL Injection, In filtration, Port scan and Botnet. (citation).

CICIDS 2017 has all the new attacks along with real world data. As the dataset is huge and vast attacks with detailed packet analysis is available, we are going to use only DoS/DDoS attack dataset. It comprises more than 80 features where the packet data is available in detail. The feature and new information available related to attack will help in building strong IDS generating the ability to detect attacks in real time. It consists of 6,92,704 records just for 5 DoS/DDoS attack types. Specific DoS/DDoS attacks are DoS SlowHttpTest, DoS Hulk, DoS SlowLoris, DoS HeartBleed and DoS GoldenEye. These attack patterns will help in training the supervised learning algorithm for increased accuracy. Various attacks in our dataset used for training are –

1. DoS - In DoS false requests are sent to the server, in order to make the offering resources unavailable to authorized users.
2. Heart-bleed -It sends a malformed heartbeat request which is sent with small payload and large length to vulnerable party to draw out the response.
3. DoS SlowHttpTest - SlowHttpTest is a tool-based attack, it highly relies on http protocol. The request sent with a very slow transfer rate which results in keeping the resources busy and waits till the whole packet is arrived.
4. DoS slowLoris - It targets many web server resources and keep them busy and hold them open as long as possible, resulting the other users fail to avail the service.
5. DoS Goldeneye – It brings down the server to knees within 30 seconds as well as it can be used for load testing

6. DoS Hulk – It targets and tries to load the server with huge amounts of packet data. As the name suggests, it will send the server huge packet payloads.

IV. Methodology

1. Data Preprocessing

Pre-processing is an important step before fitting the dataset on classifiers. The dataset consists of some values like Infinity, Nan which are not understood by the classifier at times of training. Also, huge range gap in numeric values may lead to huge training time of a classifier. So, we have performed pre-processing before passing dataset to classifier for training.

1. Removing Nan and Infinity values.
2. Removing duplicate columns and duplicate values.
3. Normalizing numeric values – Min-Max Normalization

Specific range for normalization is selected i.e. from +25 to -25. All the higher and lower range numeric data will in range from +25 to -25 where summation of whole column numeric data will result to 1.

Min-Max Normalization formula –

$$v' = \frac{v - \mathbf{min}\, A}{\mathbf{max}\, A - \mathbf{min}\, A}(\mathbf{new}_{\mathbf{max}} A - \mathbf{new}_{\mathbf{min}} A) + new_{min} A$$

2. Feature Selection

Having more than 80 features and using them to train over classifier is tedious and heavy task. It will result in overfitting of the classifier which may degrade the performance time of classifier prediction capability. So, we have used 1 feature selection algorithm – Correlation among columns.

2.1. Correlation

It refers to find mutual relationship amongst entities. It is a useful concept as it will in predicting one entity from another. Attacks patterns are chosen based on relationship of one attribute with another. So, such relationships with correlation coefficient >= 96 and <=1 are selected for feature selection. We will use Pearson Correlation Coefficient. It will measure the linear association between continuous variables. Pearson Correlation for every set of features is calculated using formula-

$$\rho_{X,Y} = \frac{\Sigma\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\Sigma\left(X_i - \bar{X}\right)^2 \Sigma\left(Y_i - \bar{Y}\right)^2}}$$

3. Classifiers – Supervised Learning Algorithms

3.1. Decision Tree -

Decision Tree is as tree like structure used to determine a course of action. Each branch indicates decision and their possible outcomes, cost and consequences. It is represented using if-else conditional statements. The ability to separate nodes denotes user decision represents clarity and transparency as compared to other decision-making algorithms. Its comprehensive nature makes it easier to build.

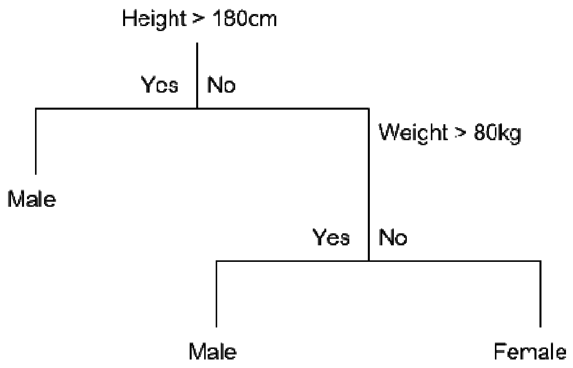


Figure 1. Decision Tree Example

Decision tree focus on relationship among different entities increasing robustness of algorithms. It is best decision-making algorithm due to quantitative analysis and statistical validation of results amongst different entities.

3.2. K-Nearest Neighbor

k-Nearest Neighbour is simplest machine learning algorithm. It is a type of lazy learning or instance-based learning method where the computation is postponed until classification. k-NN is non-parametric method that classifies object based on majority of votes from neighbour.

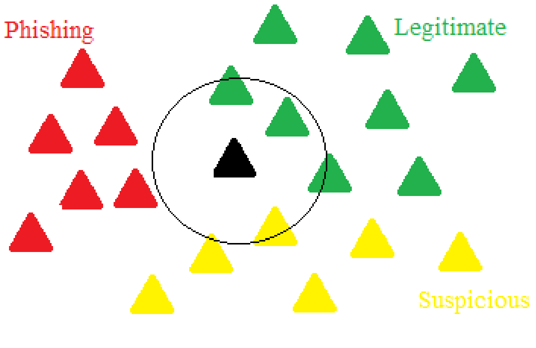


Figure 2. K-NN Example

4. Classifiers – Unsupervised Learning Algorithms

4.1. K-Means Clustering with outlier detection

K-means algorithm works on unlabelled data where the class or attack type is not known. Its goal is to find groups from data, where the number of groups is represented by K. It uses iterative approach to assign each data point to belonging group based on selected features used by algorithm. The algorithm works as follows –

1. Initialize k centroids, randomly from the dataset C1, C2...Ck.
2. For every training instance X:
   a. Calculate Euclidian distance *D (Ci, X), i = 1...k*
   b. Locate cluster Cq that is closes to X.
   c. Assign Cq to X. Compute the average of X and update the centroids.
3. Iterative till all data items are categorized into clusters and all centroids stabilize.

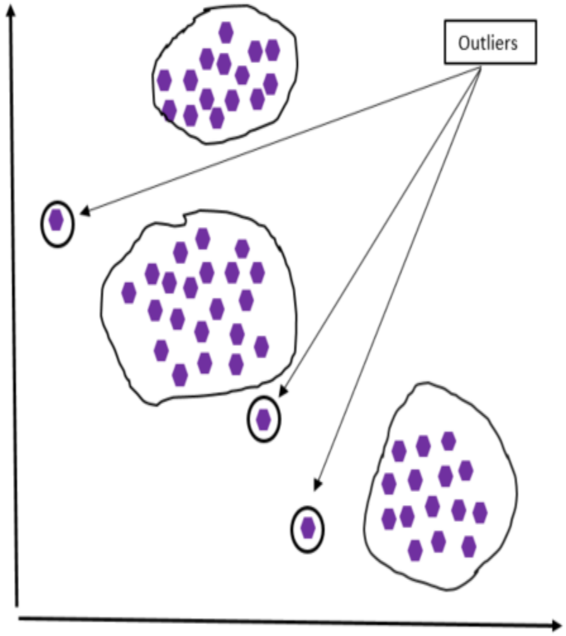


Figure 3. Clustering and Outlier Detection Example

K-means has the ability to detect and group the malicious activities that are not recognized by the supervised learning algorithms. Selecting proper features for clustering and calculating mean will easily help uncover the potential threats and identify normal behaviour. The distance between the cluster and current item will determine the class of attack. Basically, 2 clusters – anomaly and normal will be used to categorize the respective packet in real time environment.

V. CONCLUSION AND FUTURE SCOPE

By combining the anomaly-based and signature-based detection system, it is possible to benefit from both of its strengths. Hybrid and distributed approach with simple machine learning algorithms increases detection rate. Preprocessing, feature selection and extraction reflects the performance of classifiers by avoiding overfitting. Using anomaly detection, it becomes possible to detect new attacks and add them as signature to existing dataset. Adding the new and unknown attack as a signature to the existing dataset is major challenge.

In future work, we plan to introduce more accurate and intelligent classifiers that will improve the efficiency of hybrid model. Efficient network traffic analysis can also be target in future. For better feature selection Genetic Algorithm can be used to find best possible subset of features that increase accuracy of classifiers. Lastly, Sub Space Clustering can be integrated with K-means to improve the anomalous behaviour detection.

ACKNOWLEDGMENT

We would like to acknowledge and thank Prof S. R. KHONDE WHO provided expertise of domain and INSIGHTS THAT greatly helped during THE RESEARCH of this paper. We are really grateful for her constant support, valuable guidance and encouragement that helped us complete the research.

REFERENCES

[1] C. Guo, Y. Ping, N. Liu, S. Luo, “A two-level hybrid approach for intrusion detection”, Science Direct, Neurocomputing, Appl. 214, pp. 391–400, June 2016

[2] G. P. Spathoulas and S. K. Katsikas, “Reducing false positives in intrusion detection systems”, Science Direct, Computers and Security, Appl. 29, pp. 35-44, July 2009

[3] P. Casas, J. Mazel, P. Owezarski, “Unsupervised Network Intrusion Detection Systems: Detecting the Unknown without Knowledge”, Science Direct, Computer Communication, pp. 772-783, Jan 2012

[4] H. Sarvari, M M. Keikha, “Improving the Accuracy of Intrusion Detection Systems by Using the Combination of Machine Learning Approaches”, IEEE, International Conference of Soft Computing and Pattern Recognition, pp. 334-337, June 2010

[5] A. Shenfield, D. Day, A. Ayesh, “Intelligent intrusion detection systems using artificial neural networks”, Science Direct, ICT Express 4, pp. 95-99, May 2018

[6] P. Aggarwala, S. Sharma, “Analysis of KDD Dataset Attributes - Class wise For Intrusion Detection”, Science Direct, Computer Science, Appl. 57, pp. 842-851, 2015

[7] R. Ashfaq, X. Wang, J. Z. Huang, H. Abbas, Y. He, “Fuzziness based semi-supervised learning approach for intrusion detection system”, Science Direct, Information Science, Appl. 378, pp. 484-497, May 2016

[8] S. Peddabachigaria, A. Abrahamb, C. Grosanc, J. Thomasa, “Modelling intrusion detection system using hybrid intelligent systems”, Science Direct, Journal of Network and Computer Applications, Appl. 30, pp. 114-132, June 2005

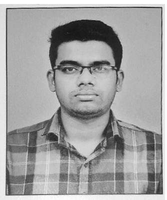
[9] S. Khonde, V. Ulagamuthalvi, “A Machine Learning Approach for Intrusion Detection using Ensemble Techniques - A survey”, International journal of scientific research in computer science, Engineering and Information Technology, Vol 3. Issue 1, ISSN - 2456-3307, pp. 328 – 338, 2018

[10] P. Rutravigneshwaran, “A Study of Intrusion Detection System using Efficient Data Mining Techniques”, International Journal of Scientific Research in Network Security and Communication, Volume-5, Issue-6, December 2017

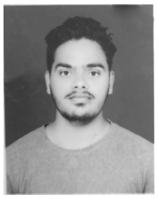
Authors Profile

A. A. Ujeniya is student currently studying in Modern Education Society’s College of Engineering Pune. He is pursuing Bachelors in Computer Engineering in 4 year program.

R. D. Pawar is student currently studying in Modern Education Society’s College of Engineering Pune. He is pursuing Bachelors in Computer Engineering in 4 year program.



S. A. Sonawane is student currently studying in Modern Education Society’s College of Engineering Pune. He is pursuing Bachelors in Computer Engineering in 4 year program.



S. B. Shingade is student currently studying in Modern Education Society’s College of Engineering Pune. He is pursuing Bachelors in Computer Engineering in 4 year program.

S. R. Khonde is Assistant Professor in Modern Education Society’s College of Engineering Pune. She is currently pursuing her Ph. D from Sathyabama Institute of Science and Technology, Chennai, India. She has published her works in 9 International Journals, 3 National Conferences and 3 International Conferences