January - February 2020
ISSN: 0193 - 4120 Page No. 8958 - 8964

Implementation on Ensemble Classification Method for Detecting Known and Unknown Web Attacks

[1]Annie Chacko, [2] Dr. A Antoni Doss, [3] Reshma Sherin Jacob
[1] Research Scholar-PTE, Dept. of CSE, Hindustan Institute of Technology and Science, Chennai, India,
[2] Associate Professor, Dept. of CSE, Hindustan Institute of Technology and Science, Chennai, India,
[3]Second Year M-Tech Student, Dept. of CSE, MBCCET, Idukki, India
[1]annievargh@gmail.com, [2] aro.antoni@gmail.com,
[3]reshmasherinjacob9048@gmail.com

Article Info
Volume 82
Page Number: 8958 - 8964
Publication Issue:
January-February 2020

Article History
Article Received*: 18 May 2019*
Revised*: 14 July 2019*
Accepted*: 22 December 2019*
Publication*: 09 February 2020*

Abstract:
The quick increment in web correspondence has been expanding generally, so there is a requirement for better security assurance. Most often the security specialists can just distinguish the known assaults. To beat the issue here proposes the data mining methods for distinguishing the unknown assaults. The procedure includes evacuating the undesirable highlights utilizing filter and wrapper method. The data mining calculations are utilized to identify obscure assaults. For finding the obscure assaults the information mining procedures like clustering, classification and their comparing calculations utilized in this strategy and afterward include extraction utilizing the filter technique. By then, the outfit procedure surveyed with the dataset, specifically the NSL-KDD dataset.

Keywords: *Anomaly Based IDS, filter methods, wrapper methods..*

I. INTRODUCTION

Network security contains the courses of action and practices got to foresee and screen unapproved get the chance to, misuse, adjustment of conceivable framework viewpoints. Network security incorporates the endorsement of access to data in a framework, which is compelled by the framework manager [1], [2]. Users select an ID and secret key or other checking data that licenses them access to data and endeavors inside their position. Frameworks can be private, for instance, inside an association, and others which might be accessible to the network. Framework security related with affiliations, adventures, and diverse sorts of foundations. It does as its title illuminates: It checks the framework, similarly as verifying and overseeing task is being done. The most generally perceived and essential technique for getting a framework resource is by assigning it another name. Despite the fact that convincing to stay away from unapproved get to, this part may disregard to check possibly hurtful substance, for instance, PC worms or Trojans.

An anomaly based intrusion ID [3], [14] structure may screen the framework like wire shark traffic and may log for survey purposes and later irregular state examination. Honeypots, principally impersonation framework accessibleresources may be sent all together as perception and early-advised gadgets, as the honeypots are not usually gotten to for authentic purposes. Such examination may likewise be utilized to fix security of the incredible structure being ensured by the honeypot. A honeypot can in like way sort out an attacker's idea far from

authority servers. A honeypot urges aggressors to contribute their time and vitality on the official server while diverting their view from the information on the genuine server. Its inspiration is furthermore to invite strikes with the objective that the attacker's systems can consider, and that information can be used to extend sort out security. A honeynet is similar to the one like honeypot [16] ordinarily.

II. IMPLEMENTATION

The paper is implemented to find out the attacks that are unknown, where unknown attacks are those attacks which is not already known on other hand their behaviours, patterns or characteristics are not previously known to the operators. There are two kinds of assaults:
Passive Attacks

Active Attacks
A passive assault is a framework in which a system is checked and all over analysed for open ports and vulnerabilities. The explanation behind existing is only to get information about the goal, and no data is changed on the objective. Passive Attacks are in watching out for or checking of transmission. The objective of the attacker is to acquire data that is being transmitted.
An active assault is a system maltreatment in which a software engineer endeavours to make changes to information on the objective. It is a sort of a strike in which a software engineer endeavours to adjust the data on the target or changing the method for the report. Some of the attacks are:-

1. DOS Attack
A DDoS happened when distinctive frameworks flood the data trade to a system to make a service unavailable to access. Such an assault is from time to time the result of different traded off structures flooding the focused on the framework with traffic. The path toward making the framework out of reach for the customer is DOS attack.

2. DNS Spoofing
DNS spoofing, similarly insinuated as DNS store hurting, is a sort of PC security hacking in which degenerate Domain Name System data is brought into the DNS resolver's reserve, making the name server return an off base outcome record, for example an IP address.

3. Phishing
Thephisher's or attackers will influence the affirmed customer to get to a site which may be a false one through messages or associations. The customer has no idea about him been using the fake site which resembles the first site yet as a rule under the control of aggressor. Phishing is a kind of social structuring that routinely used to take customer data, including login capabilities and Visa numbers. It happens when an attacker, assuming the presence of a trusted in the component, traps a harmed individual into opening an email, content, or text.

4. Buffer Overflow
Sending more data to the system that isn't typical by the structure is support flood.

5. SQL injection
Toattack data driven applications by imbuing SQL declarations into the field for execution is in fact suggested by SQL mixture. SQL implantation must undertaking a security lack of protection in an application's item, for example, when customer input is either incorrectly isolated for string exacting escape characters embedded in SQL clarifications or customer input isn't explicitly and all of a sudden executed.
For the analysis of attacks, here it considers the NSL-KDD dataset. It can be downloaded from the net. By using the data mining techniques, the important data that is needed will be mined. Information mining is the route toward discovering structures in considerable enlightening records including strategies at the intersection purpose of AI, estimations, and database systems[4], [5]. Information mining is an interdisciplinary subfield of programming designing and bits of knowledge with a general target to expel

January - February 2020
ISSN: 0193 - 4120 Page No. 8958 - 8964

information from an informational index and change the data into an intelligible structure for further use. Beside the rough examination step, it moreover incorporates database and data the board points of view, data pre-getting ready, display and deducing contemplations, multifaceted nature contemplations, post-treatment of discovered structures, portrayal, and web refreshing.

We must be familiar with the supervised and unsupervised learning. Supervised learning is the data mining task of surmising a capacity from named preparing data. The preparing information comprise of a lot of preparing models. In regulated adapting, every precedent is a couple comprising of an info object and an ideal yield esteem. A supervised learning calculation breaks down the preparation information and produces a gathered capacity, which can be utilized for mapping new precedents. An ideal situation will consider the calculation to accurately decide the class names for inconspicuous occurrences. This requires the taking in calculation to sum up from the preparation information to concealed circumstances in a reasonable manner. Unsupervised learning is the preparation of artificial intelligence (AI) calculation utilizing data that is neither arranged nor marked and enabling the calculation to follow up on that data without guidance. In unsupervised learning, an AI framework may gather unsorted data as indicated bysimilarities and contrasts despite the fact that there are no classifications given. Artificial intelligence frameworks fit for unsupervised learning are frequently connected with generative learning models, despite the fact that they may likewise utilize a recovery based methodology.

Mainly used data mining techniques [6], [17] for the project:

1. Clustering
Clustering is the way toward making a gathering of unique items into classes of comparative objects. A bunch of information articles can be treated as one group. Whereas doing cluster examination, it first segments the arrangement of information into gatherings dependent on information similarity and afterward assigns the marks to the groups. The fundamental favourable position of grouping over characterization is that, it is versatile to changes and helps single out valuable highlights that recognize distinctive gatherings.

2. Classification
A data mining technique limits the named things in collection to target groupings or classes. The objective of depiction is to precisely anticipate the genuine class for each case in the information. It uses the managed learning technique where predefined names are allotted to models by properties. For both mark and abnormal activities from the recognizable pieces of proof, portrayal estimations is being utilized. Standard or intrusion names are given to the framework traffic data accumulated in maltreatment disclosure. To learn classifiers this enlightening accumulation id used with the objective that it might be utilized for perceiving the known intrusions.

Before all the above processes, as a first step or as a pre-processing step here it considers the method called filtering.

A. FILTER METHOD
Filter strategy [10] is a champion among the most basic and once in a while used frameworks in data pre-handling for data mining. Filter techniques is characterized as utilizing some genuine property of the information so as to choose highlight utilizing the grouping calculation.
To detect the attacks here uses the NSL-KDD dataset[15], by referring this dataset the features, which are important is selected in the filter method using the Correlation BasedFeature Selection algorithm[9]. It is a just a pre-processing step, where only the relevant features is only taken. In feature selection here considers about two terms called feature reduction and feature extraction. Features contain the information about the target. Along with the filter method one more is there named as wrapper method

January - February 2020
ISSN: 0193 - 4120 Page No. 8958 - 8964

[12]. When there this more features, more will be the information and better classification power. When more relevant and redundant features comes it may lead to limited training example and limited resources. In features extraction it actually transforms the original set of feature into new subspaces, which has smaller number of dimensions. The features are selected by calculating the entropy, mutual information and symmetrical uncertainty[13]. Entropy is a measure of the uncertainty or unpredictability in a system. Information gain or alternatively called the mutual information is the measure of data that is picked up by knowing the estimation of the property, which is the entropy of the dispersion before the split minus the entropy of the circulation after it. The biggest data gain is comparable to the littlest entropy. Symmetric uncertainty uses as the proportion of connection between either two features or a component and the objective idea. In this, the co-relationship between the constant component and class feature is discovered by utilizing symmetric uncertainty measure. In the event, if the esteem is higher than the breaking point esteem, by then the component will be picked. The Correlation Based Feature Selection Algorithm [11] is as follows:

S=□
Qmax=0
for each i=1,…k, fi □ Fn-1
if exists j □ {i+1,…,k} such that |□ ij| >□ then
for each j=i+1,…..,k, i □ j, fj □ Fn-1
if Qmax < Q(Fn-1 U {fi, fj}) then S={fi, fj}
Qmax < Q(Fn-1 U {fi, fj})
endif
end for
else
if Qmax < Q (Fn-1 U {fi}) then
S={fi}
Qmax<Q(Fn-1 U {fi})
endif
endif
end for
Fn=Fn-1 U S

Where Q is quantitative criteria, Fn-1 is the initial feature set, fi and fj are features selected, □ is the parameter controlling the selection process, k is the classifier taken, Fn is the feature set with new features added.

The features selected are protocols, dst_byte, same_srv_rate, diff_srv_rate, dst_host_srv_count, count, dst_host_count dst_host_same_srv_count, dst_host_diff_srv_rate, src_byte,

dst_host_same_srv_port_rate, dst_host_srv_serror_rate, dst_host_rerror_rate, srv_count.

The features are selected or considered by taking the value of the symmetric uncertainty. The features whose SU (symmetric uncertainty) values less than the given threshold values are considered. All these features are used because these become relevant for attack detection and thereby finding the connections to the same host.

B. CLUSTERING

After the filtering procedure, next step is the clustering technique. Clustering [8] is an unsupervised learning, without using the labelled dataset; grouping is done according to the features. Clustering is viewed as an unsupervised learning strategy since we don't have the ground truth to look at the yield of the grouping calculation to the genuine marks to assess its execution. Clustering process is done using the K-Means algorithm. In clustering what actually happens is that all the features will be grouped into clusters based on their relation. It tends to be characterized as the assignment of recognizing subgroups in the information with the end goal that information focuses in a similar subgroup (group) are fundamentally the same as while information focuses in various clusters are altogether different. There will be a set of points, from that choose centroids and calculate the distance of each points to the centroids and find out which points depends on which cluster based on the distance calculated. One with the minimum distance will be put in one cluster

January - February 2020
ISSN: 0193 - 4120 Page No. 8958 - 8964

and the others will be in another cluster. Later by taking these clusters, it will again find the centroid and repeat the same procedures. K-means calculation is an iterative calculation that attempts to make the dataset into Kpre-characterized unmistakable non-covering subgroups (bunches) where every data point has a place with just a single gathering. It attempts to make the inter cluster group information focuses as comparative as would be possible whereas additionally keeping the clusters as various (far) as would be possible. It doles out information focuses to a cluster to such an extent that the total of the squared separation between the information focuses and the group's centroid (arithmetic mean of the considerable number of information indicates that have a place in that cluster) is at the base. The less variety we have inside cluster, the more homogeneous (comparable) the information focuses are inside a similar group. K-means algorithm works as follows:

1. Specify number of clusters K.
2. Initialize centroids by first rearranging the dataset and after that randomly choosing K data points for the centroids without substitution.
3. Keep emphasizing until there is no change to the centroids. i.e., task of information points to groups isn't evolving.
1. Compute the aggregate of the squared separation between information points and all centroids.
2. Assign every data point to the nearest cluster (centroid).
3. Compute the centroids for the clusters by taking the average of the all information indicates that have a place
in each group.

In clustering method the assaults is classified as normal and attacks and then they are labelled as 0's and 1's. After giving them the labels, the assaults are grouped as, all the normal labelled as 0 are together and all the attacks labelled as 1 are together. Thus, there will be clusters where the points are grouped together.

C. CLASSIFICATION

After the clustering procedure, next step is the classification technique. Classification [7] is an supervised learning, using a labelled dataset; classification is done based on the clustered information and generate a model. Supervised learning as the name demonstrates a nearness of supervisor as instructor. Fundamentally, supervised learning is a learning in which we teach or train the machine utilizing information which is very much named, that implies a few information is as of now labelled with right answer. From that point forward, machine s furnished with new arrangement of examples (data) so supervised learning calculation investigates the training data (set of preparing models) and delivers a right result from labelled information. For classification, there are two stages, testing dataset and training dataset. Training set is the one on which is trained and fit the model fundamentally to fit the parameters though testing dataset is utilized just to evaluate execution of model. Training information's yield is accessible to show whereas testing information is the unseen information for which predictions must be made. In other words, in training dataset, using the labelled dataset generation of the trained model takes place and during the testing phase, the testing process (prediction of attacks) is done in the trained model.

With the output after clustering the data will be divided into the testing dataset and training dataset. In the training dataset we will consider some attacks and in the testing dataset, there will be another set of attacks where some of them will be the same attacks present in the training dataset.

Here it detects during the testing process which all attacks satisfies the features which were selected after the filter method and at last it helps to detect whether the attack is a known or an unknown attack. The proposed system is shown in Fig.1.

For classification, here it uses the CART [7] (Classification and Regression Tree)

January - February 2020
ISSN: 0193 - 4120 Page No. 8958 - 8964

algorithm. Classification and Regression Trees or CART for short is an abbreviation introduced by Leo Breiman with allude to Decision Tree calculations that can be utilized for order or regression predictive demonstrating issues. The portrayal of the CART show is a binary tree. This is a similar double tree from algorithms and data structures(each node can have zero, a couple of small nodes). A node addresses a singular data variable (X) and a split point on that factor, tolerating the variable is numeric. The leaf nodes (in like manner called terminal node) of the tree contain a output variable (y) which is used to make a desire. Once made, a tree can be explored with another column of information following each branch with the parts until a last expectation is made. Making a binary decision tree is really a procedure of isolating up the info space. An insatiable methodology is utilized to isolate the space called recursive binary splitting. This is a numericalstrategy where every one of the qualities are arranged and distinctive split focuses are attempted and tried utilizing a cost capacity. The split with the best cost (most reduced expense since we limit cost) is chosen. All information factors and all conceivable split points are assessed and picked in an eager way based on the cost capacity. The cost function that is minimized to choose split points is the sum squared error across all training samples. The Gini cost work is utilized which gives a sign of how pure the nodes are, the place node virtue refers to how blended the training information allocated to every node is.

Fig.1: The Proposed System Design

IV. CONCLUSION

In the past, there have been so many studies that are done to study the intrusion detection system. An anomaly-based IDS using ensemble classification approach for detecting the unknown attacks has been used. The process involves removing the unwanted features using filter procedure. The data mining algorithms are used to identify unknown attacks. For finding the unknown attacks the data mining techniques like classification, clustering and their corresponding algorithms used in this technique and then feature extraction using the filter method. The evaluation of this has been done using the NSL-KDD dataset. So here consider the attacks where these attacks classified, as usual, i.e., known attacks and also unknown Then the feature selection is made, as mentioned above there are training phase and testing phase, so the data are divided into these two phases. In the training phase then the features are reduced using the filter method and now in hand have the reduced features and clustering is done. Again from the reduced elements of the training dataset, thecharacteristics are matched with the testing dataset and based on that the classification. At last the resulted output is detected as for whether the attack to be known or unknown attacks. The signature of the anonymous attacks will be updated in the database. Accuracy and detection rate will be increased by this approach.

V. REFERENCES

1. Ruzaina Khan, Muhammed Hasan, "Network Threats And Security Measure: A Review" in International Journal of Advanced Research in Computer Science and Software Engineering, Vol.8 Issue 8, 2017.
2. Komal Gandhi, “Network Security Problems And Security Attacks," 2016.
3. Rafath Samrin, D. Vasumati, “A Review On Anomaly Based Network Intrusion Detection System" in International Conference on Communication, Computer and Optimization Technique, Vol. 3, 2017.
4. Ankit Naik, S W Ahmad, “Data Mining Technology For Efficient Network Security Management" in International Journal of Computer Science Trends and Technology, Vol. 3, May-June 2015.
5. Mohammed Babiker, Yasar Hosan, "Web Application Attack Detection And Forensics: A Survey” in International Symposium on Digital Forensic and Security, 2018.
6. Sachin S Patil, Deepak Kapgate, “A Review On Detection Of Web-Based Attacks Using Data Mining Techniques" in International Journal of Advanced Research in Computer Science and

January - February 2020
ISSN: 0193 - 4120 Page No. 8958 - 8964

Software Engineering, Vol.3, Issue 3, 2013
7. Kajal Rai, Ajay Guleria, " Decision Tree Based Algorithm For Intrusion Detection" in International Journal of Advanced Networking and Technology, Vol.7, Issue 4,2016.
8. Noppol Thangsupachai, Supachai Wanapu, "Clustering Datasets With Apriori-Based Algorithm And Concurrent Processing" in International MultiConference of Engineers and Computer Scientist, Vol.1, 2011.
9. S Vishalaski, V Radha, "A Literature Review Of Feature Selection Techniques" in International Journal of Computational Intelligence and Computing Research, 2014
10. Noelia Sanchez, Amparo Alonso, "Filter Methods For Feature Selection: A Comparative Study" in International Conferenceon Intelligent Data Engineering and Automated Learning,Vol.4881, 2011
11. K. Michalak, H Kwasnik, "Correlation-Based Feature Selection Method" in International Journal of Bio-Inspired Computation, Vol. 2, Issue 5, 2010
12. Ron Kohavi, George H John, "Wrappers For Feature Subset Selection" in Journal of Artificial Intelligence, Vol. 97, 2010
13. Saurabh Mukarjee, Neelam Sharma, "Intrusion Detection Using Naive Bayes With Feature Reduction" in Elsevier, Vol. 4, 2012
14. Masaaki Sato, Hirofumi Yamaki, "Unknown Attack Detection Using Feature Extraction From Anomaly Based Ids Alerts" in International Symposium on Application and Internet, 2012
15. Hee-so Chae, Sanh Hyung Choi, "Feature Selection For Intrusion Detection Using NSL-KDD" in Recent Advances in Computer Science, 2012
16. Motahareh Dehghan, Babak Sadeghiyan Improving honey for automatic generation of attack signatures in International Journal of Intelligent Information System, 2014
17. Hu Zheng Bing, Shirochin V. P, "Data Mining Approaches for Signatures Search In Network Intrusion Detection" in IEEE
18. Workshop on Intelligent Data Acquisition and Computer Systems, 2015