International Conference on New Horizons in Science Engineering Technology (NHSET-2018)
on
13th April 2018
Organized by
Department of Electrical And Electronics Engineering, New Horizon College of Engineering, Bangalore, Karnataka, India
© 2018 IJSRCSEIT | Volume 4 | Issue 5 | ISSN : 2456-3307

Improved K-Mean Algorithm for Detection Rate in Intrusion Detection System with AI

Susheel Kumar Tiwari[1], Dr. Chandikaditya Kumawat[2], Dr. Manish Shrivastava[3]

[1]Research Scholar,Mewar University,Chittorgarh,Rajasthan, India

[2]Professor, Department of CSE, Mewar University,Chittorgarh, Rajasthan, India

[3]Professor & Head,LNCT Bhopal Affiliated to RGPV Bhopal, Madhya Pradesh, India

ABSTRACT

Intrusion location framework is a need of the present data security area. It assumes an imperative job in discovery of odd activity in a system and cautions the system chairmen to oversee such movement. The work displayed in this proposal is an endeavor to identify such movement peculiarities in the systems by creating and dissecting the activity stream information This IDS exhibited in this postulation actualizes the k-implies approach of information digging for interruption discovery and the exception recognition approach utilizing neighborhood exception factor to distinguish the irregularities present in the rush hour gridlock stream. The k-implies approach utilizes bunching systems to amass the movement stream information into typical and odd groups. The calculation is a cycle system and necessitates that the quantity of bunches, k, be given from the earlier. This choice of k esteem itself is an issue and once in a while it is difficult to anticipate before the quantity of bunches that would be there in information. This issue is settled by utilizing a metaheuristic strategy, .the man-made consciousness approach are utilized in k-mean calculation which make alterations that expansion the estimation of their target work at every single step and give better recognition rate in interruption identification framework.

Keywords : Intrusion Detection system, K-Mean, Data Mining

I. INTRODUCTION

Distinguishing the interruption in the framework assumes a vital job in the system or PC framework. Recognizing the interruption is the strategy of reviewing the activities which happens in a system and examining those activities for trace of occasions that are mischievous activities or certain danger of mischievous activities of approaches in security of the framework, allowable utilize strategies, or different practices of security. At the point when a gatecrasher endeavors to access into frameworks basic data or executes any task which is unlawful, we call this occasion as an Intrusion. There can be outer or inward interlopers which rely upon the dimension of approval. Strategies in interruption can be bugs in programming misuse's or setups of frameworks, splitting the secret key, sniffing movement which isn't anchored, or investigating the conventions which are exact and discovering configuration abandons in it. An Intrusion Detection System (IDS) is a sort of framework programming for recognizing the interruptions and advising them decisively to the best possible expert. IDSs are normally restricted to the working framework on which they work and it is the most critical instrument for the full execution of security approach of association's data that shows articulation of an association by portraying practices and guidelines for giving security, taking care of different sorts of interruptions. There are two sorts by which we can characterize IDS they are: Anomaly recognition and Misuse identification. Inconsistency Detection: Anomaly recognition makes a typical practices database and any progressions from the ordinary activity are experienced cautioning is incited that there is a plausibility of interruption in the system. Abuse Detection: Misuse Detection framework keeps up the recently characterized assault designs in the database and in the event that equivalent sort of potential outcomes happens in a system, it is delegated assault.

1.1 Attacks

Assaults are ordered into following sorts:

1. Denials-of Service: In this sort of assault, Attacker keeps veritable clients from utilizing an administration. Regularly surges the unfortunate casualties framework or system with ping messages making them hard to utilize them. (e.g. Ping of Death, smurf, SYNflood and so forth.)

2. Probing or Surveillance: Attacks have the point of getting applicable or essential data of the presence setup of a PC framework or system. Port Scans or clearing of a given IP address go by and large fall in this category.(e.g. holy person, ports sob, mscan, nmap and so forth.)

3. User-to-Root: In this injured individual machine get to is acquired by Attacker locally and gets administrator level benefits of the unfortunate casualties machine.(e.g. Perl, xterm, and so forth.)

4. Remote-to-Local (R2L): In such sort of assault, there will be unfortunate casualties machines on which aggressor won't approach rights and thus will endeavor to acquire the entrance. (e.g. word reference, guess_password, phf, sendmail, xsnoop, and so forth.)

II. RELATED WORK

In this area, related writing about catching live system movement information and age of pre-handled informational collections from crude system activity will be examined. Likewise the different sorts of information mining calculations which can be connected on datasets will likewise be talked about.

Praveen P. Naik, Prashantha S. J [2], the fundamental objective of their methodology was to distinguish interruption utilizing information mining procedures. The info informational index was KDD container design informational index. The dataset was then isolated into two sections i.e. preparing information and testing information. At that point K-implies grouping calculation was connected into ksubsets on preparing information where k is the quantity of bunches that are required for bunching. After the age of K-group neuro-fluffy (FNN) was given as contribution to every k-bunch. The yield of neuro-fluffy was given as contribution to Support Vector Machine (SVM) arrangement. At last after order i.e. utilizing SVM decided if there was interruption or not in the given informational collection.

Amine Boukhtouta, Nour-EddineLakhdari [3], the principle objective of this methodology was ID of malevolent activity at system level. In this methodology first they gathered vindictive activity by making utilization of dynamic malware investigation device and spared the crude system movement as pcap records. In the following stage they gathered non-pernicious movement from a DARPA [8] dataset and stamped it as typical. Both pcap documents i.e. malignant and typical were joined together and were exposed to include extraction. Later different machine learning calculations were connected with the goal that different classifiers could be built which can identify malevolent activity at system level. This methodology made critical upgrades when contrasted with different methodologies by catching encoded movement.

David Mudzingwa and Rajeev Agrawal [4], the ascent in the break in security of PC frameworks and

PC systems has prompted the ascent in the quantity of security devices that investigate in ensuring against these ruptures. Among these devices are interruption discovery and counteractive action frameworks (IDPS).This paper tries to offer a down to earth way to deal with assess both equipment and programming based IDPS utilizing freely accessible open source devices Tomahawk and Wireshark.

Mrs. Ghatge Dipali D [5], the primary objective of her methodology was to recognize interruption in the system by making utilization of different information mining procedures, for example, Decision tree and K-implies calculations. She made utilization of DAPRA informational index which was utilized both for preparing and also testing. A short time later DAPRA dataset was preprocessed with the goal that applicable data can be extricated from crude system information. In the following stage in the wake of preprocessing K-means and choice tree calculations were connected on preprocessed information with the end goal to recognize odd and ordinary movement.

T. Subbhulakshmi1, S. G. Keerthiga2 and R. Dharini3 [6] concocted Intelligent Multi Layered Attack Classification System (IMLACS) which helped in distinguishing and grouping interruptions with brilliant exactness in arrangement. The proposed technique caught the bundles that are transmitting through the system and extricated applicable qualities from those caught parcels. Subsequently significant properties were spared in a record. This record was utilized as contribution for help vector machine (SVM) which is a parallel classifier. SVM yield channels the records that are recognized as an assault and this is given as contribution to neural systems on which preparing and testing is finished. Neural Networks yield was given as contribution to Fuzzy derivation framework (FIS).According to the tenets experienced in FIS; it identified the kind of assault. In this methodology Real time dataset was utilized as a contribution for different characterization strategies.

S. Prayla Shyry [7], the principle objective of this methodology was distinguish bots in the system by utilizing K-implies bunching calculations. Bots are only PC frameworks or servers that are in charge of propelling different kinds of assaults, for example, disavowal of administration assaults, sending spam messages, figure secret phrase assault and so on. This strategy made utilization of botminer calculation. Right off the bat organize movement was caught utilizing system catching apparatuses .After catching pertinent data was extricated from caught information, for example, source IP, goal IP, source port, goal port, conventions and so forth. Just the bundles which started the association i.e. syn (synchronization) signal empowered and parcels which were recognized were sifted and qualities, for example, stream every hour, bytes every hour, bundles every hour and so forth were ascertained. After that mean and fluctuation were computed and K-implies bunching calculation was connected. In conclusion, subsequent to grouping it sifted assaulted bundles and typical parcels.

III. DATA MINING. WHAT IS IT?

Information mining (DM)[5], likewise called Knowledge-Discovery and Data Mining, is the procedure of consequently scanning vast volumes of information for examples utilizing affiliation rules. It is a genuinely ongoing point in software engineering however uses numerous more established computational systems from insights, data recovery, machine learning and example acknowledgment.

i. Here are a couple of particular things that information mining may add to an interruption identification venture:
ii. Remove typical movement from alert information to enable experts to center around genuine assaults
iii. Identify false alert generators and "terrible" sensor marks
iv. Find odd action that reveals a genuine assault

v. Identify long, progressing designs (distinctive IP address, same movement) To achieve these undertakings, information excavators utilize at least one of the accompanying procedures:

vi. Data rundown with insights, including discovering anomalies
vii. Visualization: introducing a graphical outline of the information
viii. Clustering of the information into characteristic classes
ix. Association rule revelation: characterizing typical action and empowering the disclosure of abnormalities
x. Classification: anticipating the class to which a specific record has a place
xi. Learning is the data which can be changed over into information about recorded examples and future patterns. The Knowledge Discovery in Database (KDD) process is for the most part characterized with the stages

1. Determination
2. Pre-preparing
3. Change
4. Information Mining
5. Understanding/Evaluation[6]

Information mining is a procedure to separate data and learning from an expansive number of deficient, boisterous, fluffy and irregular information. It is a reasonable method for separating designs, which speaks to mining totally put away in vast informational collections and spotlights on issues identifying with their attainability, value, adequacy and adaptability.

Information mining comprises of five noteworthy components

i. Extract, change, and load exchange information onto the information distribution center framework.
ii. Store and deal with the information in a multidimensional database framework.
iii. Provide information access to business investigators and data innovation experts.
iv. Analyze the information by application programming.
v. Present the information in a valuable arrangement, for example, a chart or table.

IV. SYSTEM ARCHITECTURE

The system architecture is as shown in figure 4.1. The IDS developed in this thesis consists of the following modules

· **Packet capturing module** : The packets arriving from the internet are captured by this module in real time and are stored in a pcap file for further analysis. The captured packets are of any protocol as and when they are arriving.

· **Packet reading module** : This module opens the pcap file and reads the packets contained in it. The packets are grouped according to their protocols in a file. This module writes the TCP, UDP and ICMP packets to an external file for further analysis.

· **Flow exporter module**: This module groups the packets into the flows. The features from the packets are extracted and read by this module based on which a flow record is generated. A flow generally consists of the following five parameters.

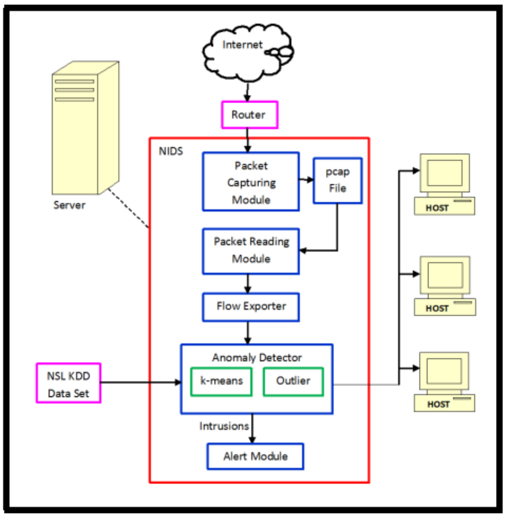
- Source IP.
- Destination IP.
- Protocol.
- Source port.
- Destination por

If there is a deviation in any of these flow values then a new flow record is generated. However, the work presented in this thesis groups the packets in accordance with the most commonly used flow records protocol called as the NetFlow version 5. The flow exporter module, therefore, groups the packets into flows according to the following fields.

Flow record ID.

- Layer 4 protocol (TCP, UDP or ICMP).
- Source IP.
- Source port.
- Destination IP.

- Destination port.
- Total packets in flow record.
- Total bytes in flow record
- Anomaly detector module



Fig(4.1):System Architecture

This module hosts the k-means and outlier detection algorithms to detect the intrusions present in each flow record. Each flow record is passed to each of the algorithms to detect the intrusions individually. The k-means approach makes use of the NSL-KDD Dataset and pcap file captured in international competitions to learn about the different types of anomalies in the network traffic. This knowledge was used to analyze the flow data by both the approaches in this module

4.1 Alert module

Based upon the analysis done by the algorithms in the anomaly detector module on each flow record using k-means and outlier detection approach, the alert module declares each flow record as normal or anomalous individually by both the approaches

V. PROBLEM STATEMENT

The idea of this thesis is to implement an NIDS that detects the anomalies present in the traffic flow. The NIDS shall use two methods of anomaly detection; k-means method and outlier detection approach. The performance of these two methods is evaluated and comparative results shall be presented in terms of various performance metrics of intrusion detection. This chapter shall present the architecture and implementation details of the IDS developed to detect anomalies in the traffic flow using the k-means and outlier detection approach.

VI. PERFORMANCE METRICS FOR IDS

The performance metrics for IDS are following

· **False positive rate (FPR**): The FPR is defined as the probability by which the IDS outputs an alert when the behavior of the traffic is normal. In this case, the IDS incorrectly gives an alert as output. The FPR can be expressed mathematically as

$$FPR = \frac{FP}{numberofnegatives}$$

False negative rate (FNR): The FNR is defined as the probability by which the IDS does not outputs an alert when the behavior of the traffic is anomalous. In this case, the IDS incorrectly does not gives an alert as output. The FNR can be expressed mathematically as

$$FPR = \frac{FN}{numberofpositives}$$

VII. PROPOSED APPROACH

This k-means algorithm aims at minimizing a squared error function is given in Equation for the objective function.

$$J = \sum_{i=1}^{k}\sum_{i=1}^{n}\|x_i(j) - c_j\|^2$$

Where $\|x_j(j) - c_j\|^2$ is a chosen distance measure between a data point xj (j) and the cluster centre cj is an indicator of the distance of the n data points from their respective cluster centers. One of the main disadvantages to K-Mean algorithm is that it requires the number of clusters as an input to the algorithm. The algorithm is incapable of determining the appropriate number of clusters and depends upon the user to identify this in beforehand. For example, if you had a group of people that were easily clustered based upon gender while calling the k-means algorithm with k=3 would force the people into three clusters and when k=2 would provide a more natural fit. Likewise, if a group of individuals were easily clustered based upon home state and you called the k-means algorithm with k=20 then the results might be too generalized to be effective.

But finding the value of i that best suits of data is very difficult. Hence we moved on to hill climbing. Hill climbing is good for finding a local optimum (a good solution that lies relatively near the initial solution) but it is not guaranteed to find the best possible solution (global optimum) out of all possible solutions (search space) which can be overcome by using steepest ascent Modified Hill climbing finds globally optimal solution. The relative simplicity of the algorithm makes it a popular first choice amongst optimizing algorithms and it is widely used in artificial intelligence, in order to reach a good state from a start state. Selection of next node and starting node can be varied to give a list of related algorithms. This can often produce a better result than other algorithms when the amount of time available to perform a search is limited, such as with real-time systems. Artificial Intelligence approach based Hill climbing algorithm attempts to maximize (or minimize) a target function $f(x)$ where x is a vector of continuous and / or discrete values. In each iteration, hill climbing will adjust a single element in x and determine whether the change improves the value of $f(x)$. Then, x is said to be globally optimal

Artificial Intelligence approach based Hill Climbing aided k-means Algorithm steps are shown bellow.

Input: randk - random value of kΔk - A random move in cluster

Output: k - Number of clusters Pseudo code: Modified Hill Climbing Algorithm

```
do
l1: iter =true;
 ksolved ← randk;
l2: newsolution ← ksolved + Δk;
 if (f (newsolution) < f (ksolved ) then
solution ← newsolution;
ksolved ← solution; k←ksolved;
 if (algorithm converged and globally optimum) then
 output k;
 iter = false;
else goto l2 ;
else goto l1 ;
while (iter);
```

Input: E= { e1, e2…en } - Set of entities to be clustered

k - number of cluster from Modified Hill Climbing Algorithm MaxIters - Limit of iterations

Output: C= {c1, c2…cn } - Set of clustered centroids

L= {l (e) e= {1, 2…n} - Set of cluster labels of E

Pseudo code:

Modified Hill Climbing aided k-means Algorithm

```
for each ci ϵ C
do ci ← ej ϵ E (E.g. random selection);
end
for each ei ϵ E do
L (ei) ← argmin Distance (ei, ci)j ϵ {1,…, k};
 end changed ← false;
 iter ← 0; repeat
 for each ci ϵ C do
Update cluster (ci);
End
 for each ei ϵ E do
 minDist ← argminDistance (ei ,cj) jϵ {1…k};
 if minDist ≠ l (ei) then;
l(ei) ← minDist;
changed ← true;
end
end
iter ← iter+1;
 until changed=true and iter ≤ MaxIters;
```

In the above algorithm is the best K value is obtained by modified hill climbing and this value is utilized in k– means algorithm in order to form effective clusters with uniform cluster density. The following section deals with performance evaluation of implemented system.

VIII. CONCLUSIONS

The Intrusion Detection System causes individuals and association to identify the assaults, programmers, and their logging data and report these I arrangement to the proprietor of the PC framework. The Intrusion Detection System not just recognizes the assault on the PC framework, it likewise decides issues with current security arrangements. In the time of Internet, numerous Internet related assaults trade off the security of PC framework. Accordingly, we should give security from these kinds of assaults and interruption recognition framework comes in help for this. We can develop Intrusion Detection Systems on different stages. One such stage is information mining. In this printed material, we give an effective Intrusion Detection System utilizing bunching procedure of Data Mining.

IX. REFERENCE

[1]. A.M Chandrasekhar, K.Raghuveer," Intrusion detection technique by using K-means, Fuzzy Neural Network and SVM classifiers", proceedings of ICCCI,pp1-7,2013(IEEE).

[2]. Praveen P Naik, Prashantha S J."An Approach for Building Intrusion Detection System by Using Data Mining Techniques "International Journal of Emerging Engineering Research and Technology (IJEERT) Volume 2, Issue 2, May 2014, PP 112-118.

[3]. Amine Boukhtouta, Nour-Eddine Lakhdari," Towards Fingerprinting Malicious Traffic", The 4th International Conference on Ambient Systems, Networks and Technologies (Science Direct).

[4]. David Mudzingwa and Rajeev Agrawal." Evaluating Intrusion Detection and Prevention Systems Using Tomahawk and Wireshark", Department of Electronics, Computer and Information Technology North Carolina A&T State University, Greensboro, NC, USA.

[5]. Mrs. GhatgeDipali D. – "Network Traffic Intrusion Detection System using Decision Tree & K-Means Clustering Algorithm" (IJETTCS) International Journal of Emerging Trends & Technology in Computer Science, Volume 2, Issue 5, September – October 2013.

[6]. T. Subbhulakshmi1, S. G. Keerthiga2 and R. Dharini3 – "Real-Time Intelligent Multilayer Attack Classification System" ICTACT Journal On Soft Computing, January 2014, Volume: 04, Issue: 02.

[7]. S. PraylaShyry, Efficient Identification of Bots by KMeans Clustering.

[8]. S. Terry, B. Chow, 1999 DARPA Intrusion Detection Evaluation Data Set,http://www.ll.mit. edu/mission/ communications/cyber/CSTcorpora/ideval/data/1999data .html.