International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No. (02), January 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

ARTIFICAL INTELLIGECE IMPROVING K-MEAN ALGORITHM FOR BETTER DETECTION RATE IN INTRUSION DETECTION SYSTEM

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

[1]*PhD Research Scholar Mewar University, Chittorgarh, Rajasthan (India)*

[2]*Professor & Head (CSE) L.N.C.T Bhopal, Affiliated to R.G.P.V Bhopal, M.P. (India)*

ABSTRACT

Intrusion detection system is a necessity of today's information security domain. It plays a vital role in detection of anomalous traffic in a network and alerts the network administrators to manage such traffic. The work presented in this thesis is an attempt to detect such traffic anomalies in the networks by generating and analyzing the traffic flow data This IDS presented in this thesis implements the k-means approach of data mining for intrusion detection and the outlier detection approach using neighborhood outlier factor to detect the anomalies present in the traffic flow. The k-means approach uses clustering mechanisms to group the traffic flow data into normal and anomalous clusters. The algorithm is an iteration procedure and requires that the number of clusters, k, be given a priori. This selection of k value itself is an issue and sometimes it is hard to predict before the number of clusters that would be there in data. This problem is resolved by using a metaheuristic method, .the artificial intelligence approach are used in k-mean algorithm which make modifications that increase the value of their objective function at each and every step and provide better detection rate in intrusion detection system.

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

I. INTRODUCTION

Detecting the intrusion in the system plays an important role in the network or computer system. Detecting the intrusion is the methodology of auditing the actions which occurs in a network and inspecting those actions for hint of events that are misbehaviors or certain risk of misbehaviors of policies in security of the system, admissible use policies, or various practices of security. When an intruder tries to access into systems critical information or executes any operation which is illegal, we call this event as an Intrusion. There can be external or internal intruders which depend on the level of authorization. Techniques in intrusion can be bugs in software exploitation's or configurations of systems, cracking the password, sniffing traffic which is not secured, or exploring the protocols which are precise and finding design defects in it. An Intrusion Detection System (IDS) is a type of system software for identifying the intrusions and informing them precisely to the proper authority. IDSs are usually limited to the operating system on which they operate and it is the most important tool for the full implementation of security policy of organization's information that displays statement of an organization by describing practices and rules for providing security, handling various types of intrusions. There are two

types by which we can classify IDS they are: Anomaly detection and Misuse detection. Anomaly Detection: Anomaly detection makes a normal behaviors database and any changes from the normal action are encountered warning is prompted that there is a possibility of intrusion in the network. Misuse Detection: Misuse Detection system maintains the previously defined attack patterns in the database and if same kind of possibilities occurs in a network then it is classified as attack.

1.1 Attacks

Attacks are classified into following types:

1. Denials-of Service: In this kind of attack, Attacker prevents genuine users from using a service. Typically floods the victims system or network with ping messages making them difficult to use them. (e.g. Ping of Death, smurf, SYNflood etc.)

2. Probing or Surveillance: Attacks have the aim of obtaining relevant or important information of the existence configuration of a computer system or network. Port Scans or sweeping of a given IP address range generally fall in this category.(e.g. saint, ports weep, mscan, nmap etc.)

3. User-to-Root: In this victim machine access is obtained by Attacker locally and obtains admin level privileges of the victims machine.(e.g. Perl, xterm, etc.)

4. Remote-to-Local (R2L): In such kind of attack, there will be victims machines on which attacker will not have access rights and hence will try to obtain the access. (e.g. dictionary, guess_password, phf, sendmail, xsnoop, etc.)

II. RELATED WORK

In this section, related literature about capturing live network traffic data and generation of pre-processed data sets from raw network traffic will be discussed. Also the various types of data mining algorithms which can be applied on datasets will also be discussed.

Praveen P. Naik, Prashantha S. J [2], the main goal of their approach was to detect intrusion using data mining techniques. The input data set was KDD cup format data set. The dataset was then divided into two parts i.e. training data and testing data. Then K-means clustering algorithm was applied into ksubsets on training data where k is the number of clusters that are required for clustering. After the generation of K-cluster neuro-fuzzy (FNN) was given as input to each k-cluster. The output of neuro-fuzzy was given as input to Support Vector Machine (SVM) classification. Finally after classification i.e. using SVM helped to determine whether there was intrusion or not in the given data set.

Amine Boukhtouta, Nour-EddineLakhdari [3], the main goal of this approach was identification of malicious traffic at network level. In this approach first they collected malicious traffic by making use of dynamic malware analysis tool and saved the raw network traffic in the form of pcap files. In the next step they collected non-malicious traffic from a DARPA [8] dataset and marked it as normal. Both pcap files i.e. malicious and normal were combined together and were subjected to feature extraction. Later various machine learning algorithms were applied so that various classifiers could be constructed which has the ability to detect malicious traffic at network level. This approach made significant improvements as compared to other approaches by capturing encrypted traffic.

David Mudzingwa and Rajeev Agrawal [4], the rise in the breach in security of computer systems and computer networks has led to the rise in the number of security tools that explore in protecting against these breaches. Among these tools are intrusion detection and prevention systems (IDPS).This paper seeks to offer a practical approach to evaluate both hardware and software based IDPS using publicly available open source tools Tomahawk and Wireshark.

Mrs. Ghatge Dipali D [5], the main goal of her approach was to detect intrusion in the network by making use of various data mining techniques such as Decision tree and K-means algorithms. She made use of DAPRA data set which was used both for training as well as testing. Afterwards DAPRA dataset was preprocessed so that relevant information can be extracted from raw network data. In the next step after preprocessing K-means and decision tree algorithms were applied on preprocessed data in order to identify anomalous and normal traffic.

T. Subbhulakshmi1, S. G. Keerthiga2 and R. Dharini3 [6] came up with Intelligent Multi Layered Attack Classification System (IMLACS) which helped in detecting and classifying intrusions with excellent accuracy in classification. The proposed method captured the packets that are transmitting through the network and extracted relevant attributes from those captured packets. Afterwards relevant attributes were saved in a file. This file was used as input for support vector machine (SVM) which is a binary classifier. SVM output filters the records that are detected as an attack and this is given as input to neural networks on which training and testing is done. Neural Networks output was given as input to Fuzzy inference system (FIS).According to the rules encountered in FIS; it detected the type of attack. In this approach Real time dataset was used as an input for various classification techniques.

S. Prayla Shyry [7], the main goal of this approach was identify bots in the network by using K-means clustering algorithms. Bots are nothing but computer systems or servers that are responsible for launching various types of attacks such as denial of service attacks, sending spam emails, guess password attack etc. This method made use of botminer algorithm. Firstly network traffic was captured using network capturing tools .After capturing relevant information was extracted from captured data such as source IP, destination IP, source port, destination port, protocols etc. Only the packets which initiated the connection i.e. syn (synchronization) flag enabled and packets which were acknowledged were filtered and values such as flow per hour, bytes per hour, packets per hour etc. were calculated. After that mean and variance were calculated and K-means clustering algorithm was applied. Lastly, after clustering it filtered attacked packets and normal packets.

III. DATA MINING. WHAT IS IT?

Data mining (DM)[5], also called Knowledge-Discovery and Data Mining, is the process of automatically searching large volumes of data for patterns using association rules. It is a fairly recent topic in computer science but utilizes many older computational techniques from statistics, information retrieval, machine learning and pattern recognition.

Here are a few specific things that data mining might contribute to an intrusion detection project:

- Remove normal activity from alarm data to allow analysts to focus on real attacks
- Identify false alarm generators and ”bad” sensor signatures
- Find anomalous activity that uncovers a real attack

- • Identify long, ongoing patterns (different IP address, same activity) To accomplish these tasks, data miners employ one or more of the following techniques:
- Data summarization with statistics, including finding outliers
- Visualization: presenting a graphical summary of the data
- Clustering of the data into natural categories
- Association rule discovery: defining normal activity and enabling the discovery of anomalies
- Classification: predicting the category to which a particular record belongs

Knowledge is the information which can be converted into knowledge about historical patterns and future trends. The Knowledge Discovery in Database (KDD) process is generally defined with the stages

1. Selection
2. Pre-processing
3. Transformation
4. Data Mining
5. Interpretation/Evaluation[6]

Data mining is a process to extract information and knowledge from a large number of incomplete, noisy, fuzzy and random data. It is a suitable way of extracting patterns, which represents mining completely stored in large data sets and focuses on issues relating to their feasibility, usefulness, effectiveness and scalability.

Data mining consists of five major elements

1. Extract, transform, and load transaction data onto the data warehouse system.
2. Store and manage the data in a multidimensional database system.
3. Provide data access to business analysts and information technology professionals.
4. Analyze the data by application software.
5. Present the data in a useful format, such as a graph or table.

3.1 Advantages of Data Mining Techniques

i. Problems with large databases may contain valuable implicit regularities that can be discovered automatically [7].
ii. Difficult-to-program applications, which are too difficult for traditional manual programming.
iii. Software applications that modify to the individual users preferences, such as modified advertising.

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.

International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No. (02), January 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

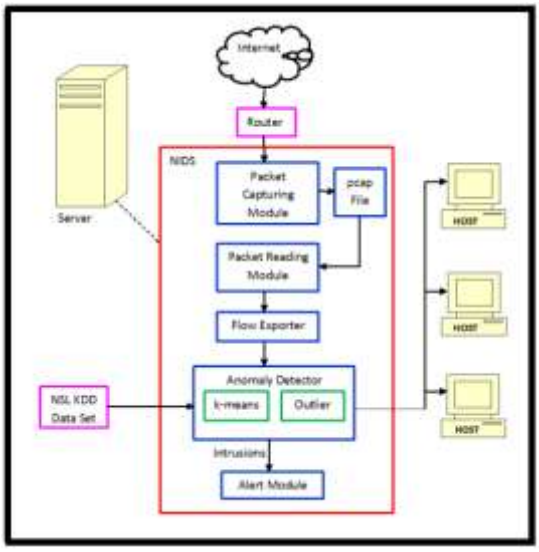
· **Flow exporter module**: This module groups the packets into the flows. The features from the packets are extracted and read by this module based on which a flow record is generated. A flow generally consists of the following five parameters.

- 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.

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

Where $\left\|x_i(j) - c_j\right\|^2$ 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 helps people and organization to detect the attacks, hackers, and their logging information and report these i formation to the owner of the computer system. The Intrusion Detection System not only identifies the attack on the computer system, it also determines problems with current security policies. In the age of Internet, many Internet related attacks compromise the security of computer system. Therefore, we must provide security from these types of attacks & intrusion detection system comes in aid for this. We can construct Intrusion Detection Systems on various platforms. One such platform is data mining. In this paper work, we provide an efficient Intrusion Detection System using clustering technique of Data Mining.

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/ communi cations/cyber/CSTcorpora/ideval/data/1999data.html.