IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-6, Issue-10, Oct 2018 E-ISSN: 2347-2693

# Intrusion Detection and Prevention System to Increase the Detection Rate Using Data Mining Technique

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

[1] Department of CSE, Mewar University, Chittorgarh, Rajasthan,India
[2]Department of CSE, Mewar University,Chittorgarh,Rajasthan,India
[3]LNCT Bhopal Affiliated to RGPV Bhopal (M.P) India

**Corresponding Author: sushiltiwari24@yahoo.co.in*

**Available online at: www.ijcseonline.org**

Accepted: 17/Oct/2018, Published: 31/Oct/2018

***Abstract-*** Intrusion Detection Systems are used to monitor computer system for sign of security violations over network or cloud environment. On detection of such sign triggers of IDSs is to report them to generate the alerts. These alerts are presented to a human analyst or user who evaluates the alerts and initiates an adequate response. In Practice, IDSs have been observed to trigger thousands of alerts per day, most of which are mistakenly triggered by begin events such as false positive. This makes it extremely difficult for the analyst to correctly identify alerts related to attack such as a true positive. Recently Data Mining methods have gained importance in addressing network or cloud security issues, including network intrusion detection and cloud Intrusion detection systems, these systems aim to identify attacks with a high detection rate and a low false alarm rate. Consequently, Unsupervised Learning methods have been given a closer look for network and cloud intrusion detection. We present unsupervised based Clustering Technique and compare with traditional centroid-based clustering algorithms for intrusion detection. These techniques are applied to the KDD Cup98 data set .In addition; a Comparative analysis shows the advantage of proposed approach over Traditional clustering-based Methods over in identifying new or unseen attack. Experimental result show that A.I based Hill Climbing aided k-means Clustering algorithm improves the detection rate in IDS than K-Mean algorithm and achieved 92% detection rate in IDS System.

***Keywords-*** Intrusion Detection, AI, Clustering

## I. INTRODUCTION

The very fast growth of development of internet and World Wide Web has changed the scenario of internet and networking technologies. This rapid growth of internet and network technologies has also increased the number of users and amount of data which travel in network. In another hand the chances of data loss, hacking and intrusion is also increased. In order to this the demand of network security techniques has also increased to keep the data and network resources secure. To provide security to data and resources various techniques has been proposed one of them is Intrusion Detection System. An intrusion detection system (IDS) is a component of the network security environment. Its main task is to identify between suspicious or intrusive activity and normal behavior. The aim of intrusion detection is to build a system which itself watch network activity and detect such intrusive attacks. Once an attack is recognized, the system administrator is informed who takes adequate action to deal with the intrusion.

The IDS works with two main approaches one is misuse detection and another is anomaly detection. Misuse detection approach is based on stored signature pattern. In this a database contains signatures and if data does not match with pre stored data it is counted as intrusion. Anomaly detection approach based on behavior pattern if data traffics behavior varies with a normal user behavior then it is counted as intruders. IDS is host-based (HIDS), network based (NIDS) or a combination of both types that is Hybrid Intrusion Detection System. HIDS usually observes logs and system calls on a single system, in another hand a NIDS typically observes traffic and data flows and Network data packets on a network segment, and that's why it checks number of hosts in a very short events. Generally, one use to work with very large amount of network data, because of this it is very hard and tiresome work to classify the records manually in order to detect a correct intrusion. Labeled data can be obtained by simulating intrusions, but this will be limited only for the set of known attacks. That's why new types of attacks which occur in future cannot be handled, if those were not part of the training data. Sometimes with manual classification, we can only able to identifying the previously known (at classification time) types of attacks, that's why we restrict our detection system to identify only those types of attacks. To solve these deficiencies, we need a technique to detect intrusions when our training data is unlabeled, as well as for detecting new and un-known types of attacks. A method that offers reliability in this task is

 

anomaly detection. Anomaly detection finds anomalies in the data (i.e. data instances in the data that deviate from normal or regular ones). It also makes us able to detect new types of intrusions, because these new types will, by assumption, be deviations from the normal network usage. It is very difficult and sometimes not possible, to detect malicious users when misfeasor is a person who uses is authorized user to use the network and who uses the network in a legitimate way. For example, there is probably not a very reliable way to know whether someone who appropriately logged into a system and work as the intended user of that specific system, or at the situation if the password was stolen.

Under all these assumptions and consideration we design a system which created clusters from its input data, then it itself form labelled clusters as containing either normal or attacks type of data instances, and finally used these clusters to classify network data instances as either normal or intrusive. Both the training and testing was done using 10% KDDCup'99 dataset [2], which is a very popular and most frequently used intrusion attack dataset. Most of the clustering techniques assume a well-defined differentiation between the clusters so that each pattern can only belong to one cluster at a time. This supposition can neglect the natural ability of object which existing in numbers of clusters. For this reason and with the aid of fuzzy reasoning, fuzzy clustering can be employed to overcome the deficiencies. The membership of a pattern in a given cluster can vary between 0 and 1. In this model a data object belongs to the clusters where it has the highest membership value.

The main goal of intrusion detection is to detect unauthorized use, misuse and abuse of computer systems by both system insiders and external intruders. Among automated intrusion detection systems, a particular system for network intrusion detection, known as a network-based intrusion detection system (IDS), monitors any number of hosts on a network by scrutinizing the audit trails of multiple hosts and network traffic. It is usually comprised of two main components: an anomaly detector and a misuse detector [1][2]. The anomaly detector establishes the profiles of normal activities of users, systems, system resources, network traffic and/or services and detects intrusions by identifying significant deviations from the normal behavior patterns observed from profiles. The misuse detector defines suspicious misuse signatures based on known system vulnerabilities and a security policy. This component probes whether these misuse signatures are present or not in the auditing trails. This paper proposes the use of negative selection and nicking of artificial immune system for developing an effective network-based IDS. An overall artificial immune model for network intrusion detection presented in consists of three different evolutionary stages: negative selection, clonal selection, and gene library evolution. Among these stages, the first stage, negative selection, is investigated in this paper. We present a more efficient implementation of negative selection using a niching feature of artificial immune systems [9]

## II. LITERATURE SURVEY

A lot of research works have been carried out in the literature for intrusion detection and some of them have motivated us to take up this research. Brief reviews of some of those recent significant researches are presented below:

**Tich Phu oc Tran** have applied Machine Learning techniques to solve Intrusion Detection problems within computer networks. Due to complex and dynamic nature of computer networks and hacking techniques, identifying malicious activities remains a challenging task for security experts, that is, defense systems that were currently available suffer from low detection capability and high number of false alarms.

**Ye Yuan et** proposed a method of evidence assignment in combination with Dempster-Shafer theory to identify network attack data. In this method, extracted features were identified by a multi-generalized regression neural network classifier, which determined the basic probability assignment.

**Snehal A** proposed the decision tree based algorithm to build multiclass intrusion detection system. Support Vector Machines was the classifiers which were initially designed for binary classification.

**Shun J and Malki H. A.** presented a neural network-based intrusion detection method for the internet-based attacks on a computer network.

**Muna Mhammad T. Jawhar and Monica Mehrotra** presented an intrusion detection model based on hybrid fuzzy logic and neural network. The key idea was to take advantage of different classification abilities of fuzzy logic and neural network for intrusion detection system. The model had capability to identify an attack, to distinguish one attack from another i.e. classifying attack, and the most vital, to detect new attacks with high detection rate and low false negative.

**Neal, and Hunt and Dasgupta, Cao, and Yang** who successfully applied their AISs to recognition and classification tasks. They showed that their IS-inspired models were flexible, noise-tolerant and generalized their classification well. It has also been shown that it is possible to perform these tasks effectively with resource limited AISs (Timmis and Neal 2008).

**Hu Zhengbing1 and et al** proposed an algorithm to use the known signature to find the signature of the related attack quickly. They used nine different-sized databases,

 

**Pohsiang Tsai et al.** suggested a Machine Learning (ML) framework in which various types of intrusions would be detected with different classifiers, containing different attribute selections and learning algorithms. Appropriate voting techniques were used to combine the outputs of these classifiers

**Aida Hu Zhengbing** proposed an algorithm to use the known signature to find the signature of the related attack quickly. They used nine different-sized databases.

**Amit Kumar Choudhary** proposed a neural network approach to improve the alert throughput of a network and making it attack prohibitive using IDS. For evolving and testing intrusion the KDD CUP 99 dataset were used.

**Stefano Zanero** proposed a novel architecture which implements a network-based anomaly detection system using unsupervised learning algorithms. They described how the pattern recognition features of a Self Organizing Map algorithm can be used for Intrusion Detection purposes on the payload of TCP network World Journal of Science and Technology 2012, 2(3):127-133 131 packets.

## III. PROBLEM IDENTIFICATION

The main drawback of traditional methods is that they cannot detect unknown intrusion. Even if a new pattern of the attacks were discovered, this new pattern would have to be manually updated into system. It is also capable of identifying new attacks to some degree of resemblance to the learned ones, the neural networks are widely considered as an efficient approach to adaptively classify patterns [11], but their high computation intensity and the long training cycles greatly hinder their applications, especially for the intrusion detection problem, where the amount of related data is very important.

## IV. PROPOSED APPROACH

We propose Artificial Intelligence based clustering algorithm for network intrusion detection. 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 €C

do ci ← ej €E (E.g. random selection);

end

for each ei €E do

L (ei) ← argmin Distance (ei, ci)j €{1,…, k};

end changed ← false;

iter ← 0; repeat

for each ci €C do

Update cluster (ci);

End

for each ei €E do

minDist ← argminDistance (ei ,cj) j€ {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

## V. CONCLUSION

This thesis has given an overview of artificial intelligence that used with clustering technique. It seems that neural networks are the most popular selection for this kind of AI with a good reason. Wide variety of choices for a neural network type makes it possible to select a type that works in a given application. Other AI types have also been proved to be suitable for IDS use. AIs seem to have the needed accuracy for IDS use. Configuring AI-based IDS is easier than con- figuring traditional IDS. This decreases deployment costs which are an important factor for companies. Because of this, they could more easily test different easy-to-deploy IDSs to see which of them is most the secure and requires the least monitoring in their network. Traditionally k means clustering algorithm has been used in the area but due to the robustness and finding the means only it is not possible to apply this algorithm directly in the intrusion data as it is collection of heterogamous data. Therefore several mixed type clustering approaches are used already to detect intrusion in the IDS area, so proposed a artificial intelligence based Hill Climbing aided K-Mean algorithm provide better performance in Intrusion Detection accuracy rate and faster running time.

## VI. REFERENCES


[1] Tich Phu oc Tran, “Machine Learning and Data Mining: Introduction to Principles and Algorithms”, Horwood Publishing Limited, 2007.

[2] Ye Yuan, "Mining Audit Data to Build Intrusion Detection Models," Proc. Fourth International Conference Knowledge Discovery and Data Mining pp. 66-72, 1999\

[3] Snehal A, "The Research of Intrusion Detection Based on Support vector machine", Proceedings of the 2008 International Conference on Wavelet Analysis and Pattern Recognition, Hong Kong, IEEE.2008

[4] Shun J and Malki H. A., "Network Intrusion Detection System using Neural Networks", IEEE computer society.2008.

[5] Muna Mhammad T. Jawhar and Monica Mehrotra, “A Study On Fuzzy Intrusion Detection”, In Proceedings of the Data Mining, Intrusion Detection, In formation Assurance, And Data Networks Security, SPIE, Vol. 5812, pp. 23-30, Orlando, Florida, USA, 2005

[6] Neal, and Hunt and Dasgupta, Cao, and Yang, “Anomaly Network Intrusion Detection Based on Improved AIS Technique”, Journal of Computers, Vol." “Adaptive Model Generation: An Architecture for the Deployment of Data Mining-Based Intrusion Detection Systems, Applications of Data Mining in Computer Security”, Kluwer Academic Publishers, Boston, MA, pp. 154-191, 2002

[7] Pohsiang Tsai,’A novel intrusion detection system based on hierarchical clustering and support vector machines", Expert Systems with Applications, Vol: 38, No: 1, pp: 306-313, 2011.

[8] Aida Hu Zhengbing, ”Approaches and machine learning Techniques for Intrusion Detection Systems”, Vol. 9, No. 12, pp. 181-186, 2009.

[9] Amit Kumar Choudhary,”An Effective Approach to Network Intrusion Detection System using Neural network technique”, International Journal of Computer Applications, Vol.1, No.3, pp.26–32, February 2010


 