International Journal For Technological Research In Engineering *ISSN (Online): 2347 - 4718*
International Conference on Emerging Technologies in Engineering, Biomedical, Medical and Science (ETEBMS - 16)

U2R ATTACK DETECTION USING MACHINE LEARNING

Purushottam R.Patil[1], Dr. Yogesh Sharma[2], Dr.Manali Kshirsagar[3]
[1]Ph.D. Research Scholar, FE&T JNU, Jodhpur, India
[2]Professor (Maths), FE&T JNU, Jodhpur, India
[3]Vice President (Academy), ADCC Infocad, Nagpur, India

Abstract: ***The performance measurement criteria of Intrusion Detection System's (IDS) include True Positive rate, False Alarm rate, and Detection rate when meets 1 in the scale of 0 to 1 said to be efficient and accurate. Most of the IDS's implementation results grasp up to 0.99 for above factors when tested for DDoS, Probe, R2L, U2R attack. In This paper we propose a system which reaches to 100 % Detection rate when tested for U2R attack. The dataset has chosen MIT Lincoln Lab's KDD'99 10% dataset which is benchmark for the same.***

Index Terms: Intrusion detection System (IDS), detection rate, True Positive rate, False alarm rate.

I. INTRODUCTION

In Computer Network when an of attempts to compromise a computer or a computer network resource security is regarded as an intrusion. The security services like data integrity, Confidentiality, validation. Intrusion detection Systems (IDS). Used to build up the system security and increase its confrontation to internal and outside attacks. Generally, the main task of IDS is to detect an intrusion and, if necessary or possible, to undertake some measures to eliminate future intrusions. In this paper, a User to Resource (U2RID) System is presented. The system is developed using the Clustering technique, K-means, Evolutionary algorithm Genetic algorithm and feed-forward Neural Network (NN). In the following section, a brief introduction to IDSs and K-Means, GA and Neural Network concepts is given. In section 3, previous works related to IDS using K-means, GA and NN have been discussed. U2RID framework and related improvements are explained in section 4. Finally, implementation results of our experiments are shown in section 5.

II. INTRUSION DETCTION SYSTEM

Intrusion detection collects the runtime assembly of data from system operation, and the subsequent analysis of the data; the data can be audit or web logs generated by an operating system, packets sniffed taken from a network, or reports from instrumented programs, which could be applications such as a DBMS. We can categorize IDSs with many different criteria. IDSs can be separated into network-based IDS and host-based IDS by the systems they monitor. IDSs that monitor network backbones and look for attack scenarios are called network-based IDSs, whereas those that operate on hosts defend and monitor the operating and file systems for signs of intrusion are called host-based IDSs.

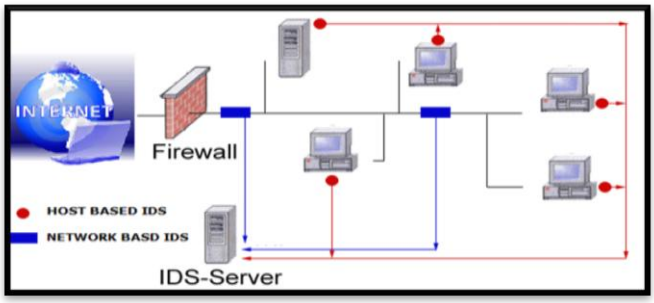

Figure 1. Intrusion Detection System

III. IMPLEMENTATION TECHNIQUES

A. K-means Clustering

K-means is one of the most popular methods used to solve the clustering problems phases is to define k centroids, for each cluster. The next phase is to take each point, measures distance is generally considered to determine the distance between data points and the centroids. When all the points are included in some clusters, the first step is completed and an early alliance is done. At this point we need to recalculate the new centroids, as the inclusion of new points may lead to a change in the cluster centroids. Oncewe find k new centroids, a new binding is to be created between the same data points and the adjacent new centroids, generating a loop. As a result of this loop, the K centroids may change their position in a step by step manner. ultimately, a situation will be reached where the centroids do not move anymore [4].

Algorithm:
Input:
D={d1,d2,…dn} // sets of n data items.
K // number of preferred clusters
Output:
A set of k clusters.
Steps:
1. Choose k data items from D as initial centroids;
2. Repeat the process of selecting the items
Assign the each item to the cluster which has the nearest and the fitting centroids; Calculate the new mean value for each cluster; Repeat the process until the criteria is satisfied.

B. Genetic Algorithm

GA are examples of evolutionary computing methods and are optimization technique of computer algorithm. GA is inspired by Dawin's theory about evolution. Algorithm is

started with a solution space represented by chromosomes called population [4]. In GA, a population of strings encodes the candidates solution to an optimization problem which involves towards better solutions. Solutions are represented in binary as strings of 0's and 1's, so the output are generated randomly or individually. In this process multiple independently are selected and it is evaluated automatically [4].

C. Basic's of GA

1. [START] Generates random items of each n Chromosomes i.e population which is suitable solutions of the problem.
2. [FITNESS] Evaluates the efficiency of function f(x) of each chromosomes x in the string.
3. [NEW POPULATION] new substring are created by repeating following steps until the new substring is created.
    1. [Selection] Selecting these two substrings from the given strings according to their nearest relationships of the given string. (Fitness value is good, better chances of selection)
    2. [Crossover] Probability cross over the parents to form a new offspring. If no cross over was performed, children is an exact copy of parent.
    3. [Mutation] A mutation probability creates a new substring for the children presented in the chromosomes.
    4. [Replace] New generated substring will be processed to run the algorithm.
    5. [Test] If the test condition is satisfied, STOP and return the best solution in current string.
    6. [Loop] Continues if the situation is not satisfied. then the process will Go to step2.

Selection of String, Grouping the String, Interrelation of the string [4]. GA is used for the purpose of finding the fixed K number of the cluster, where GA's is executed first to give initial values of k-means to start with rather than choosing random ones and expected to minimize the number of iterations that K-means needs in order to converge the local minima.

D. Principles of GA

1) Encode the data as the binary string.
2) Randomly generate the substring for the given string. This one includes a genetic illustration for the group of solutions.
3) Knowing the fitness of the substring value for each string. Directly it will depend on the distance to the optimum value of the string.
4) Selection of the string from the substring shares the data's fitness value.
5) Genomes crossover and mutations has been Processed for each values.
6) And then start again from the point3.

GA's application into clustering an initial population of random cluster is set. The k cluster centers encoded in each chromosome are initialized to K randomly chosen points from the data set. This process is repeated from each of the P chromosomes in the population. At each generation, each individual is evaluated by fitness value. New individuals are created using two main genetic operators are crossover and mutations. At the beginning of a run of a GA large populations of random chromosomes are created. Each one, when decoded will represent a different result to the problem [4].

E. Neural Network

It is a set of interconnected nodes designed to imitate the functioning of the human brain. Each node has a weighted connection to several other nodes in neighboring layers. The individual nodes take the input received from connected nodes and use the weights together with a simple function to compute output values. Neural networks can be constructed for supervised or unsupervised learning. The user specifies the number of hidden layers as well as the number of nodes within a specific hidden layer. It mostly depend on the application, the output layer of the neural network may enclose one or several nodes. The Multilayer Perceptions (MLP) neural networks have been very booming in a variety of applications and producing more accurate results than other existing computational learning models. It is Capable of approximating to arbitrary accuracy, any continuous function as long as they include enough hidden units. Means that such models can form any categorization decision boundary in feature space and thus act as non-linear discriminate function [5].

F. Implementation steps

The operational concept of the system flow is as shown in figure 2.

1) It is a method for intrusion detection using network traffic anomaly detection approach
2) Apply K-means clustering algorithm to form clusters of intrusion and normal sets of data
3) Apply Genetic Algorithm to clusters to formulate rule from the apparent traits of normal and abnormal clusters

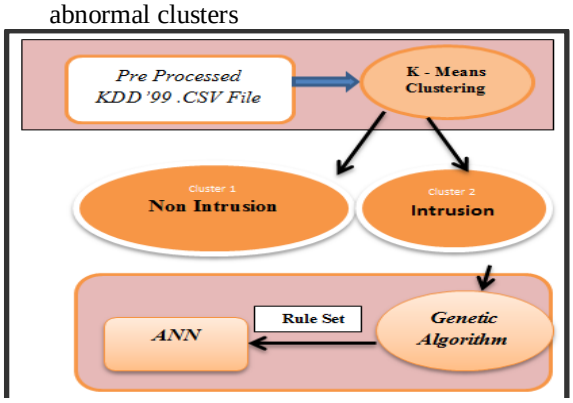

Figure 2.Proposed System Methodology

4) Train and test the artificial neural network using the knowledge gained through GA
5) Use trained ANN for detecting abnormal packets similar to the ones given during learning sessions and for an artificial intelligence that detects anomalies not presented during learning.
6) Verification and validation of proposed method using different network traffic data sets.

IV. RESULTS AND EVALUATION

The evaluation of IDS was originated by the US DARPA in 1998 and has been the most inclusive scientific study known for comparing the performance of different IDSs. The MIT Lincoln Laboratory synthesize the network traffic with its data sets DARPA 1998 and DARPA 1999. The performance of IDS can be evaluated by choosing these widely available data sets. IDSs can be configured and tuned in a variety of ways in order to reduce the false positive rate and to maximize the detection rate. However, there is a trade-off between these two metrics for any system and hence these measurements are used to form the Receiver Operating Characteristic (ROC) curves. It plots the DR Vs FP. If the IDS alarms very often on every suspicious packet, the false alarm rate as well as the detection rate will increase. On the other hand, if the IDS generates alarms only after sufficient facts is available i.e., lower false alarms, the detection rate will undergo but with an increased alarm confidence. IDS can be operated at any given point on the ROC curve. The best or optimal operating point for an IDS, given a specific network, is determined by factors like the cost of a false alarm, the value of a proper detection and the prior probabilities of normal and attack traffic. The ROC curve conveys necessary information when analysing and comparing IDSs [5]. Figure 3 is an ROC graph plotted with each point identify the status of particular IDS in terms of the DR and the FP. The packed region to the top left as seen in the graph can be identified as that due to the recent systems. Thus the environment in which most of the IDSs of modern times operate requires very lowfalse alarm rates i.e. much lower than the 0.1% selected by DARPA for useful detection. The overall accuracy is 100 % when measured as AUC of ROC as shown in Table1. The results are as shown in Figure 4.

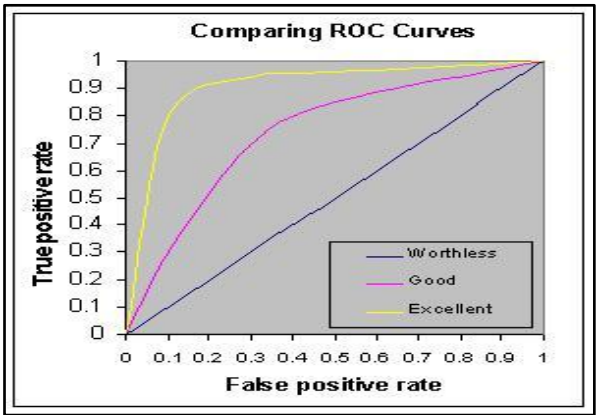

Figure 3. Comparative ROC Curve

Table 1. Experimental Results

Attack Type	Category	No of Records	AUC of ROC
U2R	Buffer Overflow	30	1
	Load Module	9	1
	Rootkit	3	1
	Pearl	10	1

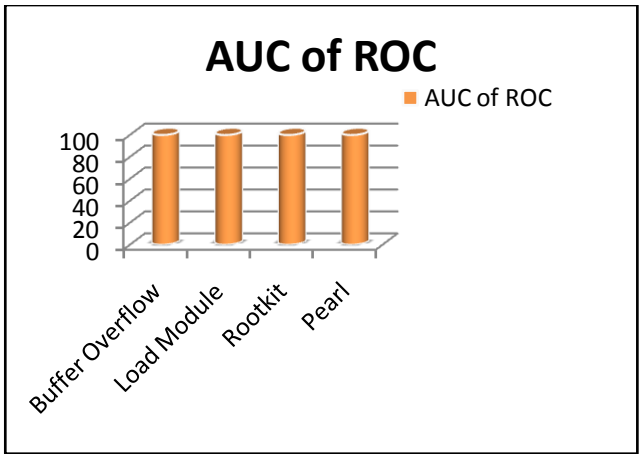

Figure 4. Results and Categories

REFERENCES

[1] Tao Song, "Formal Reasoning about Intrusion Detection Systems",Computer Science,March 2007.
[2] Ciza Thomas," Performance Enhancement of Intrusion Detection Systems using Advances in Sensor Fusion", Supercomputer Education and Research Centre IISc, Bangalore, 2009
[3] Preeti Aggarwal, Sudhir Kumar Sharma,"Analysis of KDD Dataset Attributes-class wise For Intrusion Detection", Elsevier.2015.
[4] Dr. A.V. Senthil Kumar,S.Mythili,"Parallel Implementation of Genetic Algorithmusing K-Means Clustering",Int. J. Advanced Networking and Applications, Volume:03 Issue:06 Pages:1450-1455 (2012) ISSN : 0975-0290.
[5] ShikhaAgrawal, JitendraAgrawal," Survey on Anomaly Detection using Data Mining Techniques", 19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems Procedia Computer Science 60 (2015) 708 – 713.
[6] Leonard Portnoy, Data mining Lab,Department of Computer science, Columbia University.

Purushottam Rohidas Patil obtained his Bachelor's degree in Computer Engineering in 2000 from North Maharashtra University, Jalgaon, M.S, India. Then he obtained his M.E. CSE in 2009. From Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, M.S,India .and Currently, he is a Ph.D Research Scholar at Faculty of Engineering and Technology, Jodhpur National University, Jodhpur, RJ, India. His current research interests are Data Mining, Network Security, Human Computer Interaction. He is Members of Professional bodies like LMISTE, CSI, IAENG, CSTA-ACM.

Dr.Yogesh Sharma received his B.Sc. Degree in 1995, M.Sc.(Mathematics) degree in 1997,and Ph.D. (Mathematics) in 2001 from Jai Narayan Vyas University, Jodhpur, RJ, India. His research interest includes Fractional Calculus, Differential Operators, Matrix Variable, Lie Theory, operation Research, He is currently working as Dean Faculty Of Computer Application and Professor & Head, Applied Science, Faculty of Engineering and Technology, Jodhpur National University, Jodhpur,(RJ), India. He is Members of Professional bodies like Vijnana Parished Anusandhan Patrika, Allahabad, Indian Academy of Mathematics, Indore, Rajasthan Ganita Sandesh, Society of Special function, International Scientific Committee USA, and reviewer 8 Mathematical Society. Dr. Sharma has published 9 books,41 research papers at Reputed International and National Journals and attended and participated 9 national and international Conferences. He is approved research guide at Career Point University, Jodhpur National University, JJTU, Jhunjunu, Rajasthan, India.

Dr. Manali Makarand Kshirsagar, Ph.D. (Computer Science), M.E.(Computer Science & Engineering), B.E. (Computer Technology) and Diploma (Computer Technology), M.B.A.(Fin. & Marketing). Her Specialization at the PG and Ph.D. level are Data Warehousing, Data Mining, Business Intelligence and Bioinformatics. Other areas of interest are Advanced O.S., Wireless Sensor Networks, Business Analytics .She has Total 23 years of experience in Academics and Industry. Her Job Profile is Education Leader, full Professor at UG and PG level, Ph.D. Supervisor, Head of the Department, Dean (Students Activities).Dr. Manali Kshirsagar currently held the position as Vice President(Academy) ADCC Infocad Ltd. Nagpur, India. She is Members of professional societies and positions held in local chapters, Fellow of Institution of Engineers(India), Life Member of Indian Society for Technical Education(New Delhi), Chair of CSI Nagpur Chapter, Elected Member of IE(I) Nagpur Local Center. She has Above 50 technical and research publications in International Journals, International and National conferences. Guiding 08 Ph.D. students at present.

