An Adopted Hybrid Approach for Encroachment Catching By Combining Neuro Fuzzy Clustering and SVM On KDD-Cup-Dataset

[1]Priyanka Bera
[1st] Research Scholar
[1,]Computer Engineering Department,
[1]Darshan Engineering College, Rajkot, India

***Abstract* - With the impending era of internet, the network security has become the key foundation for lot of financial and business web applications. Intrusion detection is one of the looms to resolve the problem of network security. Imperfectness of intrusion detection systems (IDS) has given an opportunity for data mining to make several important contributions to the field of intrusion detection. In recent years, many researchers are using data mining techniques for building IDS. Here, we propose a new approach by utilizing data mining techniques such as neuro-fuzzy and radial basis support vector machine (SVM) for helping IDS to attain higher detection rate. The proposed technique has four major steps: primarily, k-means clustering is used to generate different training subsets. Then, based on the obtained training subsets, different neuro-fuzzy models are trained. Subsequently, a vector for SVM classification is formed and in the end, classification using radial SVM is performed to detect intrusion has happened or not. To illustrate the applicability and capability of the new approach, the results of experiments on KDD CUP 1999 dataset is demonstrated. Experimental results shows that our proposed new approach do better than BPNN, multiclass SVM and other well-known methods such as decision trees and Columbia model in terms of sensitivity, specificity and in particular detection accuracy.**

***Index Terms* - : KDD, Intrusion Detection, SVM, Neuro Fuzzy, Encroachment Catching, and Data mining**

I. INTRODUCTION

As defined in [1], intrusion detection is the process of monitoring the events occurring in a computer network and analyzing them for signs of intrusions. It is also defined as attempts to compromise the confidentiality, integrity, availability, or to bypass the security mechanisms of a computer network. Anomaly intrusion detection systems (IDSs) aim at distinguishing an abnormal activity from an ordinary one. The current state of computer networks is vulnerable; they are prone to an increasing number of attacks. These attacks are seldom previously seen. It is very hard to detect them before subsequent damage is done. Therefore, securing such a network from unwanted malicious traffic is of prime concern. In this paper, a two-phase method for intrusion detection, called 2PID, is proposed. The Knowledge Discovery in Databases (KDD) Cup 1999 data set [2], which has been utilized extensively for development of IDSs, is used as a representative sample of data. Current anomaly detection is often associated with high false alarm with moderate accuracy and detection rates when it's unable to detect all types of attacks correctly.

To overcome this problem, Muda et al. [3] proposed a hybrid learning approach through combination of K-Means clustering and Naïve Bayes classification. They cluster all data into the corresponding group before applying a classifier for classification purpose. An experiment is carried out to evaluate the performance of the proposed approach using KDD Cup '99 dataset. Result show that the proposed approach performed better in term of accuracy, detection rate with reasonable false alarm rate. H. Om et al. [4] proposed a hybrid intrusion detection system that combines k-Means, and two classifiers: K-nearest neighbor and Naïve Bayes for anomaly detection. It consists of selecting features using an entropy based feature selection algorithm which selects the important attributes and removes the irredundant attributes. This system can detect the intrusions and further classify them into four categories: Denial of Service (DoS), U2R (User to Root), R2L (Remote to Local), and probe. The main goal is to reduce the false alarm rate of IDS.

II. RELATED WORK

Existing IDS techniques includes high false positive and false negative rate. Nadiammai et al. [5] implemented some of the clustering algorithms like k means, hierarchical and Fuzzy C Means, to analyze the detection rate over KDD CUP 99 dataset and time complexity of these algorithms. Based on evaluation result, FCM outperforms in terms of both accuracy and computational time. Y. Qing et al. [6] presented an approach to detect intrusion based on data mining frame work. In the framework, intrusion detection is thought of as clustering. The reduction algorithm is presented to cancel the redundant attribute set and obtain the optimal attribute set to form the input of the FCM. To find the reasonable initial centers easily, the advanced FCM is established, which improves the performance of intrusion detection since the traffic is large and the types of attack are various. In the illustrative example, the number of attributes is reduced greatly and the detection is in a high precision for the attacks of DoS and Probe, a low false positive rate in all types of attacks.

III. KDD CUP DATASET

The data set provided for the 1999 KDD Cup was originally prepared by MIT Lincoln labs for the 1998 Defense Advanced Research Projects Agency (DARPA) Intrusion Detection Evaluation Program, with the objective of evaluating research in intrusion detection, and it has become a benchmark data set for the evaluation of IDSs. Attacks fall into four main categories [2]:

- Denial of service (DoS), where an attacker makes some computing or memory resource too busy or too full to handle legitimate requests, thus denying legitimate users access to a machine, e.g., SYN flood;
- Remote to local (R2L), where an attacker sends packets to a machine over a network, then exploits machine's vulnerability to illegally gain local access as a user, e.g., guessing password;
- User to root (U2R), where an attacker starts out with access to a normal user account on the system and is able to exploit vulnerability to gain root access to the system, e.g., buffer overflows;
- Probing, where an attacker scans a network to gather information or find known vulnerabilities, e.g., port scanning.

The KDD Cup 1999 data set has a huge number of duplicated records as shown in Table I on the next page. This data set lies with the distribution of its five classes. The DoS attack comprises 79.24% in training and 73.90% in testing, respectively. Meanwhile, normal connection consists of 19.69% in training and 19.48% in testing, respectively. This imbalance makes it very difficult to train classifiers on the training set, and results in having extremely poor detection rates. In this paper, we use a subset of the original data set which consists of distinct records only.

duration	flag	hot	su_attempted	num_outbound_cmds	same_srv_rate
protocol_type	land	num_failed_logins	num_root	is_host_login	diff_srv_rate
service	wrong_fragment	logged_in	num_file_creations	is_guest_login	srv_count
src_bytes	urgent	num_compromised	num_shells	serror_rate	srv_serror_rate
dst_bytes	count	root_shell	num_access_files	rerror_rate	srv_rerror_rate and srv_diff_host_rate
dst_host_count	dst_host_srv_count	dst_host_same_srv_rate	dst_host_diff_srv_rate	dst_host_same_src_port_rate	dst_host_srv_diff_host_rate
dst_host_serror_rate	dst_host_srv_serror_rate	dst_host_srv_rerror_rate	class		

Table I. KDD 41 feature set

Attack	Attack Type
Dos	Back,Land,Neptune,Pod,Smurf,teardrop
Probe	Satan,Ipsweep,Nmap, Portsweep,
R2L	Guess_Password, Ftp_write,Imap, Phf,Multihop, Warezmaster, Warezclient, Spy
U2R	Buffer_overflow, Loadmodule, Rootkit

Table II. Various Attack & its attack type

IV. PROPOSED METHODS

Neuro-Fuzzy Training Module

Generally, K-means clustering results in the formation of 'K' clusters where each cluster will be a type of intrusion or the normal data. For every cluster, we have Neuro-fuzzy classifiers associated with it, i.e., there will be 5 number of Neuro fuzzy classifiers is trained with the data in the respective cluster. Neuro-fuzzy makes use of back propagation learning to find out the input membership function parameters and least mean square method to find out the consequent parameters.

The first hidden layer maps the input variable correspondingly to each membership function. In the second hidden layer, T-norm operator is used to compute the antecedents of the rules. The rules strengths are normalized in the third hidden layer and subsequently in the fourth hidden layer the consequents of the rules are found out.

SVM Vector Generation Module

Classification of the data point considering all its attributed is a very difficult task and takes much time for the processing, hence decreasing the number of attributes related with each other of the data point is of paramount importance. The main purpose of the proposed technique is to decrease the number of attributes associated with each data, so that classification can be made in a simpler and easier way. Neuro-fuzzy classifier is employed to efficiently decrease the number of attributes.

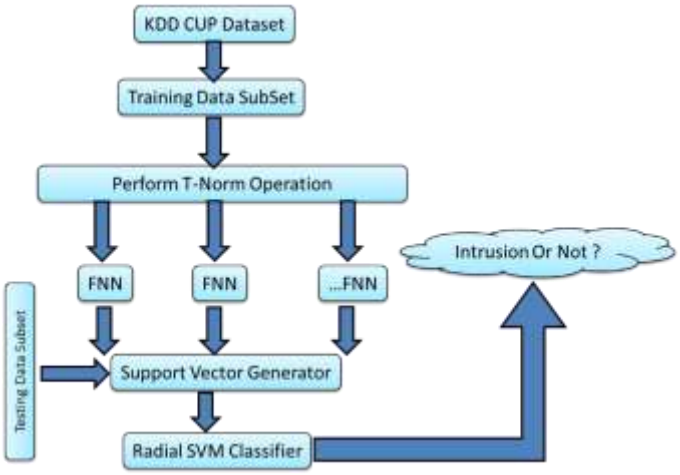

Figure 1. Proposed System Architecture

Radial SVM Classifier Module

This classifier is used as it produces better results for binary classification when compared to the other classifiers. But use of linear SVM has the disadvantages of getting less accuracy result, over fitting results and robust to noise. These short comings are effectively suppressed by the use of the radial SVM where nonlinear kernel functions are used and the resulting maximum margin hyper plane fits in a transformed feature space. In our proposed technique, nonlinear kernel functions are used and the resulting margin hyper plane fits in a transformed feature space. When the kernel used is a Gaussian radial basis function, the corresponding feature space is a Hilbert space of infinite dimensions.

V. PROPOSED ALGORITHM

Input here we use the KDD dataset. KDD dataset is containing various features for the intrusion detection system. Then, KDD dataset needed as perspective of Training Subset. In next step, it will apply Tree Normalization Operation which will generate Pruned Tree. At next, input to the Neuron Fuzzy clustering method. It will apply set of Fuzzy Rules and generate another Cluster.At last, Support vector machine use training and testing dataset for the process. Now generated training dataset use as the input to support vector machine. SVM generates the testing dataset for the next.At the end , Radial SVM classifier is applied that gives the binary value (0 | 1) as the result like false alarm. 0 indicates the no means normal data and 1 indicates the yes malicious data.

Step 1 : Input Data Di govern by attribute ai
Step 2 : For every i=1 to …..n belongs to Di
Step 3: Calculate Fuzzy Logic With Membership Function

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_i\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$

Step 4 : Training Vector Classifier Is Generated From Fuzzy C-Means Clustering
Step 5 : Input Training Vector Classifier To Support Vectors from $S_1, S_2, \ldots\ldots S_n$
Step 6 : Compute SVM Kernel Function
Step 7 : Apply To SVM Radial Classifier
Step 8 : Generate Classification Results
Step 9 : Return Encroachment Detect Or Not

V.IMPLEMENTATION & RESULTS

Implementation has been carried by using Java Programming language and on KDD dataset with taking among 10% of KDD cup dataset for experimental based. Stepwise procedure mention in proposed algorithm is been carried out to measure various output. The way each step is carried out is also mentioned in proposed architecture and as per that proper GUI with in Java is build. For system running Eclipse tool is used as an IDE. Various implementation screenshot and its report generation is also mentioned in below. At the next accuracy parameter is calculated based on TP, TN, FP, and FN.

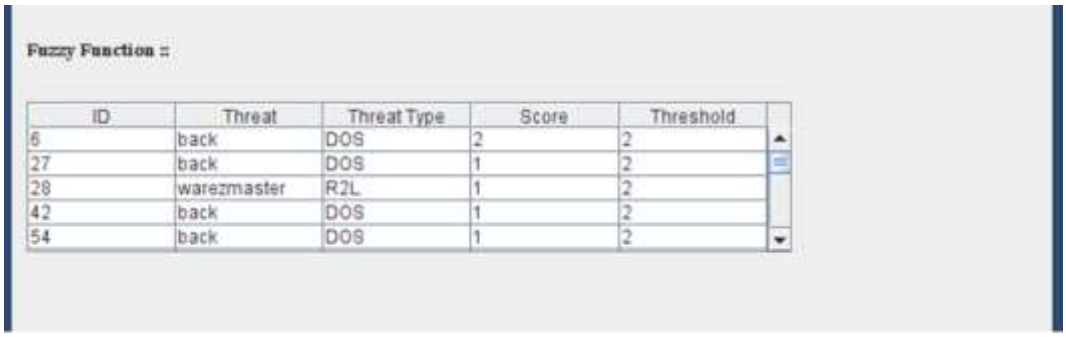

Figure 2. Fuzzy Function Output

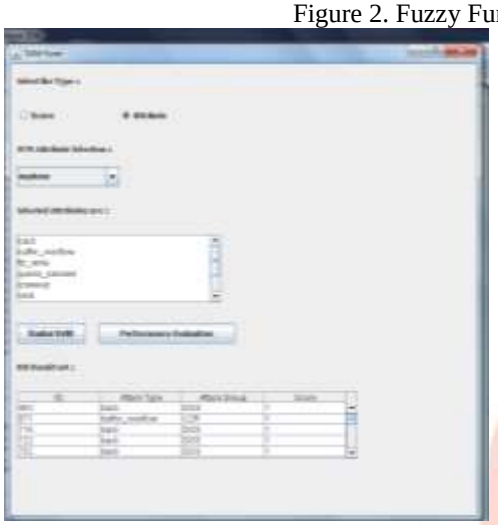

Figure 3. Radial SVM Classifier Output

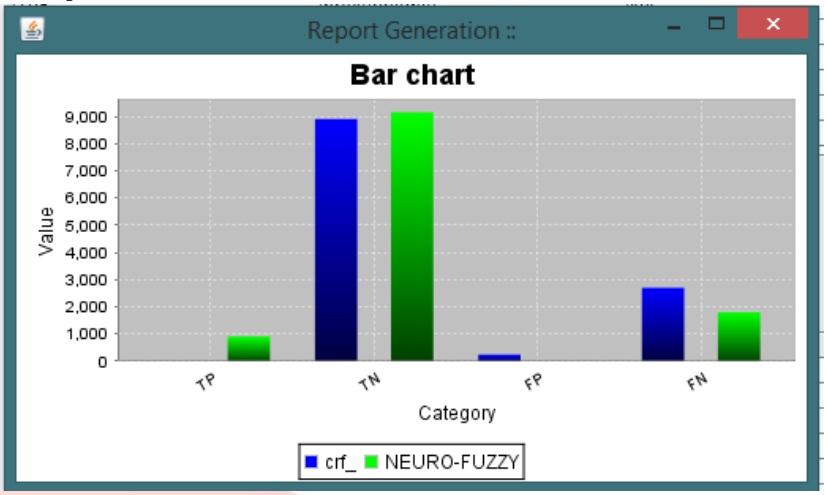

Figure 4.Report Generation

Probability of false alarms, In *TP* cases: intrusion – alarm. In *TN* cases: no intrusion – no alarm. In *FP* cases: no intrusion – alarm. In *FN* cases: intrusion – no alarm.

Total intrusions: *TP+FN*
Total no-intrusions: *FP+TN*

Where TP=True Positive, TN=True Negative, FP=False Positive, FN= False Negative.

METRIC	TYPES OF ATTACK							
	DOS		PROBE		R2L		U2R	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
True Negative	12487	12199	12487	12199	12487	12199	12487	12199
False Positive	13	301	13	301	13	301	13	301
True Positive	12500	12500	2025	1963	38	25	14	12
False Negative	0	0	29	90	1	13	7	9
Accuracy	99.95	98.80	99.71	97.31	99.89	97.51	99.84	97.52

Table III. Various Metrics & Their Value Representation for Various Attacks

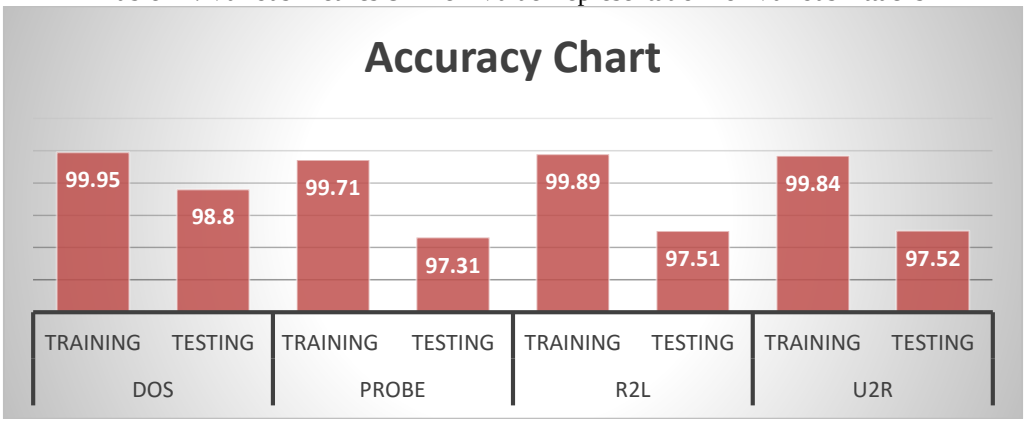

Figure 5. Accuracy Chart Representation

VI. CONCLUSION

In recent years, research on neural network methods and machine learning techniques to improve the network security by examining the behavior of the network as well as that of threats is done in the rapid force. The large volume of database is increasing rapidly resulting in gradual rise in the security attacks. The current IDS is ineffective to update the audit data rapidly it involves human interference thus reduces the performances. The paper elaborates the architecture of the Intrusion Detection System along with features of an ideal intrusion detection system. The study also describes the categorization and challenges if the IDS. In this paper we analyzed the neural network approach and the machine learning approach in overcoming the challenges of the IDS. Further there is need to design the system which will overcome the current challenges of IDS and also the system must provide a high performance in detecting the threats and security attacks.

REFERENCES

[1] R. Bace and P. Mell, "Intrusion Detection Systems," NIST Special Publications on Intrusion Detection Systems. SP 800.31, Nov. 2001.
[2]"KDDCup 1999 Dataset". [Available online]: http://kdd.ics.uci.edu/databases/kddcup1999.html/
[3] Muda Z., Yassin W., Sulaiman M. N., and Udzir N. I., (2011): "Intrusion detection based on K-Means clustering and Naive Bayes classification," in *Information Technology in Asia (CITA 11), 7th International Conference on*, Kuching, Sarawak, pp. 1-6.
[4] Om H., and Kundu A., (2012): "A hybrid system for reducing the false alarm rate of anomaly intrusion detection system," in *Recent Advances in Information Technology (RAIT), 1st International Conference on*, Dhanbad, pp. 131-136.
[5] Nadiammai G. V., and Hemalatha M., (2012): "An Evaluation of Clustering Technique over Intrusion Detection System," in *International Conference on Advances in Computing, Communications and Informatics (ICACCI'12)* Chennai, pp. 1054-1060.
[6] Ye Q., Wu X., and Huang G., (2010): "An intrusion detection approach based on data mining," in *Future Computer and Communication (ICFCC), 2nd International Conference on*, Wuhan, pp. V1-695-V1-698.