Ensemble Learning (Reptreea2deu Algorithm) Used For Intrusion Detection System

Mr. Navnit Upadhyay[1], Mr. Kailash Patidar[2]
[2] HOD Dept of CSE
[1, 2] SSSIST, Sehore (M.P.)

Abstract- *As of late information and communication technology (ICT) has turned into an imperative piece of human life. Be that as it may, ICT brings a considerable measure of cyber risks. New dangers and vulnerabilities are made to assault n/w framework. IDS plays an significant to secure the network and its major goal is to vision the network activities involuntarily to indentify the malicious attacks. Intrusion detection system also known to be (IDS) is suitable for critical component that to make safe network in the today's world. By utilizing data mining(DM) in IDS can enhance the recognition rate, dealing with the false alert rate and lessen false positive rate. Intrusion detection system (IDS) is utilized to detect these assaults. Machine learning (ML) and DM methods are broadly utilized for IDS. Present IDS algorithms result in great rate of error and less accurate to categorize various attacks. This paper manages a novel group classifier (RepTreeA2DEU) for IDS. Proposed group classifier is manufactured utilizing two surely understood calculations RepTree and A2DEUpdable. RepTree enhances exactness and diminishes the error rate. The performance of proposed ensemble classifier (RepTree+A2DEU) is analyzed on Kyoto data set. Proposed ensemble classifier outperforms A2DEUpdable and RepTree algorithms and efficiently classifies the network traffic as normal or malicious.*

Keywords- Internet, IDS, AODE, RepTree, Card, Machine Learning and WEKA.

I. INTRODUCTION

In a recent year, usage of internet is increased in all the fields. As the usage of internet is increasing in our daily life, the network security is becoming necessary in order to obtain security, integrity and confidentiality of a resource. Along with the firewalls, intrusion detection system (IDS) has become a main component of the security system. The role of IDS is to trap the hacker's presence on the network. As the large number of incidents is increasing in our daily life IDS's are used with improved techniques. IDS plays an important to secure the network and its main goal is to view the network activities automatically to indentify the malicious attacks. Intrusion detection system also known to be (IDS) is suitable for critical component that to make safe network in the today's world. By utilizing DM in IDS can enhance the identification rate, dealing with the false alert rate and decrease false positive rate. Intrusion Detection approach can identifies and deals with network those are malicious in computer and computer network resources. Keeping in mind the end goal to identifying information target IDS has been named into two classifications:

- Host-based IDS
- Network-based IDS

Host based IDS's are designed to monitor, detect and response to activity and attacks on the given host. Network based IDS's capture network traffic for their intrusion detection operations.[1]

ANNs also called as (artificial neural network) are the processing devices (algorithms or the actual hardware) which are freely modeled subsequent to the neuronal structure of mamalian cerebral cortex however that on a great deal littler scales. A large type of ANN may have hundreds or the thousands of units processor , while a mamalian brain have billions of the neurons with a consequent raise in the magnitude of their overall interaction and emergent behavior.[2]

The usage of Genetic Algorithm (GA) in networks optimization has lasted for more than a decade, minimum since the early of 2000's based on the papers we reviewed. GA is a arbitrary seeking method that is motivated from biological progression theory .It accommodates a combination of desired solution, which in this case is seen as an array of chromosome. Chromosome is a chain of genes. Each gene represents one parameter of solution; it can be of the same or of different kind. Basically, the content of every gene is a float number between 0 and 1.[3]

II. LITERATURE SURVEY

Dr. S. Dugad(2017) et al introduces about In our venture, we exhibit a cutting edge answer for send intrusion detection utilizing image processing and Support Vector Machine also called as (SVM). The key aim is to distinguish

the ships, which cross over border and protected industrial spaces. By the mechanisms of interworking of these 2 approaches, we can identify the interfering ship from continually changing the sea environment. SVM used as machine learning to instruct the system by revealing it to dissimilar environments of seashore . Subsequently, it can be utilized as an ongoing security framework at seashore territories.[7]

Om Prakash Nirankari(2017)et al presents about — Service Chaining provides opportunities for network and service providers to implement their services and policies by methods for better granularity of individual client or application. Nonetheless, the expanding number of Service Chains and middleboxes will present a bigger number of flow rules and more utilization of Ternary Content Addressable Memory (TCAM), whose limit is constrained because of high cost and power utilization. This paper proposes to compress the flow rules for benefit fastening by improving the age of Service Chain IDs that are generally utilized as a part of packet tagging procedures for the Service Chaining. Our answer 1) influences administration to chain IDs aggregatable in light of Common Forwarding Actions (CFAs) among the service chains, and 2) lessens the quantity of stream rules at each SDN change to execute a bigger number of sending activities for service chaining. The assessment comes about demonstrated that the proposed calculation can diminish up to 76% of the stream rules utilizing the arbitrarily created systems and service chains. Since the age of Service Chain ID does not meddle the other flow rule compression strategies, our calculation can likewise be utilized as a module to the next Service Chaining systems to streamline their ID generation.[8]
Anand Keshri(2016)et al presents about-Denial of Service (DoS) assaults represent a genuine danger to business organizations. DoS attacks is difficult to shield on account of various ways that hacker may strike. DoS attacks center around specific applications. DoS attack targets to make the service out of resources, so that it becomes unavailable to the legitimate users. Because of capricious conduct of hacker it is hard to recognize honest to goodness and malignant system activity. Moreover, as defence against these improve, attacks also evolve. New sort of obscure assaults proceed to strike and it is difficult to distinguish them in light of data of previous assaults. DOS attacks usually aspire websites or the services like card payment gateways, of the banks, and even the domain name servers. In this paper, we talk about DOS assaults and quickly see the distinctive counteractive action plans. At that point we examined DoS anticipation utilizing firewall and IDS and diverse ways to deal with IDS utilizing DM procedures. We utilized NSL-KDD dataset, refined adaptation of kdd'99 cup data set for applying DM algorithms and testing.[9]

H. Gharaee (2016)et al displays around an anomaly based IDS using GA and Support Vector Machine (SVM) with a new-fangled feature selection process. The new model has used a feature selection method based on Genetic with an innovation in fitness function reduce the dimension of the data, increase true positive detection and simultaneously decrease false positive detection. In addition, the computation time for training will also have a remarkable reduction. Results demonstrate that the proposed strategy can achieve high exactness and low false positive rate (FPR) all the while, however it had prior been accomplished in before thinks about separately. This study proposes a method which can achieve more stable features in comparison with other techniques. The proposed model experiment and test on KDD CUP 99 and UNSW-NB15 datasets. Numeric Results and comparison to other models have been presented.[10]

Gözde Karataş(2016)et al displays about IDS are frameworks that avoid or moderate the serving of various sorts of server information activity caused by intensive employ of systems to broaden via net. Especially as of late, due to increment in data density, requirement for these frameworks is expanded consequently unique identification calculations are being produced. In this learning, data about the dissimilar detection algorithms by means of genetic algorithm, which are made of IDS algo is given and literature search been made.[11]

Yang Liu(2016)et al exhibits about We address the undertaking of recognizing objects from video input. This fundamental issue is nearly unexplored, contrast with image based object recognition. To this end, we construct the ensuing commitments. First, we begin with two comprehensive datasets for the video based object detection. Next, we recommend Latent Bi-constraint SVM (LBSVM), for the most part intense border configuration for video-based object recognition (ObjReg). LBSVM is being based on the Structured Output SVM, however extends it to handle video data which are noisy and make sure about consistency of productivity decision in excess of the time. We apply the LBSVM that to distinguish office objects and the exhibition hall figures, and we express its advantages over picture based, set based, and extra video based ObjReg. [12]

III. INTRUSION DETECTION SYSTEM

Intrusion Detection is a key approach in the Information Security acting an imperative task of detecting dissimilar type of attacks and secure the network system. Intrusion Detection is the method of observing and analyzing the actions arising inside a computer or the network system that to recognize all the problems on security . An ID provides

3 significant functions of security; monitor, discover and react to activities those are unauthorized. IDS screens the tasks of firewalls, routers, the management servers and the basic documents to new security components. IDS have the capacity to make the security overseeing of the framework by non-master staff conceivable by methods for giving the easy to understand interface.

IDS for the most part give the accompanying administrations:

- Observing and investigating PC as well as system framework movement.
- Audit system configurations and the vulnerabilities
- Evaluate the integrity of the dangerous system and the data files
- Estimate the activities which are abnormal .

IDSs are being isolated into 2 general classifications: HIDS and second is NIDS . A HIDS requires tiny programs (or operators) to be introduced on singular frameworks to be managed. The agents examine the OS and write down the data to log files and trigger the alarms. A network based (IDS) Intrusion Detection System generally consists of network application (or sensor) with Network Interface Card (NIC) functioning in licentious mode and a separate interface management . IDS located on a network segment or the border and examine each and every traffic on that particular segment. The present pattern in interruption location is to consolidate both host based and network based information to create hybrid systems that that have more effective.[4]

IV. ALGORITHMS FOR IDS

A. Genetic Algorithm

GA is a random search technique that is inspired from biological evolution theory . It accommodates a combination of desired solution, which in this case is seen as an array of chromosome. Chromosome is a chain of genes. Each gene represents one parameter of solution; it can be of the same or of different kind. Basically, the content of every gene is a float number between 0 and 1. The gene content is generated by a random number generator. And after the gene's value is generated, we put it in a function/formula, so that it will be directed to a meaningful value that refers to a solution, for example: a formula is written that if the gene's value returns a number more than 0.5, then the parameter will become a string that states 'vertical', and if the gene's value returns a number equal or less than 0.5, then the parameter will become a string that states horizontal. Other varieties of formula can be made according to our need and objective.

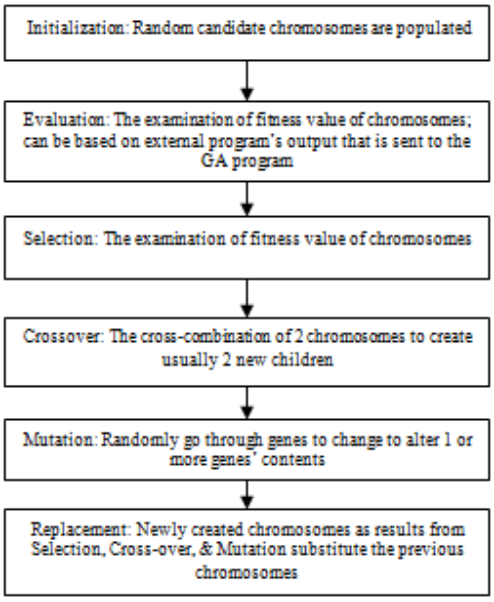

Fig: 1 The overall GA process steps are depicted in below diagram .

The above described chromosome will be bred, populated, and gone through GA process that consists of the following operators:

a) Reproduction: This operator copies the select better chromosomes in a latest population, and it used to be the main administrator actualized on population. It picks great chromosomes in a population to frame a mating pool. It plays beneficial role to inherit the best of chromosomes of the previous population, thus the quality of the new population can be maintained or improved.
b) Crossover: It cross- combines two chromosomes to generate a improved strain of chromosome. The chromosomes are recombined by exchanging information which then new chromosomes are created based on this information, hence a hybrid could appear. However, the newly created hybrid may have either reduced or improved quality, thus to preserve some good chromosomes, not all of them are used in this crossover process.
c) Mutation: The mutation occurs by changing the value of a gene in random position. The objective is to diversify the solution. [5]

B. Support vector machine

The support vector machine (SVM) generally deal with the classification of pattern which means this algo is mostly used for the classifying of dissimilar types of patterns. At the present, there is dissimilar types of patterns that is Linear and non-linear. Linear patterns are the patterns that

easily discernible or can easily be alienated in the little dimension, while nonlinear patterns are the patterns that not easily discernible or can't easily be alienated and therefore these types of patterns require to be further manipulate so they can be separated without difficulty .

Support vector machine, outstanding amongst other machine learning calculations, which was proposed in 1990 and for the most part utilized for pattern recognition. Also image detection, the speech recognition, the text classification, detection of face and the detection of faulty card , etc like a lot of paradigm has useful for the problems for classification . SVM machine learning is a fault. In the algo that, given set of training example, each are related with one of a number of category as, A model that predicts that original SVM training manufacture a scope of exam. SVM learning for normal issue, which is going for the more prominent limit factually.

SVM is based on the theory of statistical learning and the principal of structural risk minimization and include the plan of determining place of the decision limits also called as a hyper plane that produce the optimal separation of classes. Maximize the margin and in that way create the biggest possible distance among the separating hyper plane and instances on each side of it has been confirmed to decrease an superior bound on the estimated simplification error concept of the SVM also called as support vector machine on which SVM is recognized are being given as takes after:

- Separating the hyper plane.
- Most extreme edge hyper plane
- Soft edge.
- The Kernel function.[6]

V. PROPOSED MODEL

Ensemble learning is a new trend in AI and data mining, in which several weak learning algorithms are combined. Thought behind ensemble classification is to misuse the quality of powerless learning calculations to acquire a strong/effective classifier. A single IDS developed with weak learning algorithm can cover and identify limited input data and no. of attacks. Ensemble classifiers are constructed by a set of weak classifiers and decision function which combines the classification results. Majority voting is simple and efficient decision function used in many ensemble techniques.

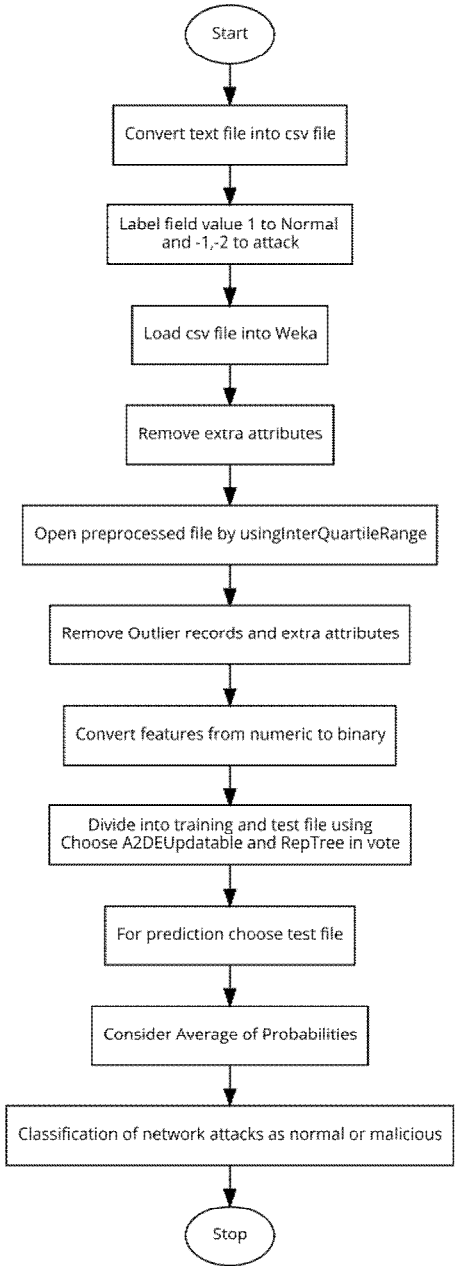

Fig.2 Proposed Model

Proposed Algorithm:

Step:1 Convert text file into csv file and Label field value 1 to Normal and -1,-2 to attack
Step:2 Load csv file into Weka –en file and select files of type .csv and select file and then press open button
Step:3 Remove extra attributes
Step:4 Open preprocessed file by using Inter Quartile Range
Step:5 Remove Outlier records and extra attributes
Step:6 Convert the features from numeric to binary Divide into training and test file using Remove Percentage
Step:7 Choose A2DEUpdatable and RepTree in vote
Step:8 For prediction choose test file

Step:9 Consider the Average of Probabilities
Step:10 Classification of network attacks as normal or malicious
Step:11 Stop

VI. RESULT & ANALYSIS

In the result analysis, the experiment of proposed work performed by using ensemble classifier. Kyoto dataset 2006 used for the investigational study of the traffic data. This dataset contains 24 features and we used only 15 features and excluded the features which are related to security analysis.

Instances: 85346
Attributes: 15
Duration_binarized
Service
SourceByte_binarized
DestinationByte_binarized
Count_binarized
Same srv rate_binarized
Serror rate_binarized
Srv serror rate_binarized
Dst host count_binarized
Dst host srv count_binarized
Dst host same src port rate_binarized
Dst host serror rate_binarized
Dst host srv serror rate_binarized
Flag
Label
Test mode: 10-fold cross-validation

Base Work:

Time taken to build model: 1.69 seconds

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.996	0.865	0.955	0.996	0.975	0.285	0.935	0.996	Attack
0.135	0.004	0.667	0.135	0.224	0.285	0.935	0.439	Normal
Weighted Avg.	0.952	0.821	0.940	0.952	0.937	0.285	0.935	0.968

=== Confusion Matrix ===

a	b	<-- classified as
80681	294	a = Attack
3783	588	b = Normal

Propose Work:

Time taken to build model: 0.38 seconds

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.998	0.735	0.962	0.998	0.979	0.463	0.940	0.997	Attack
0.265	0.002	0.854	0.265	0.405	0.463	0.940	0.512	Normal
Weighted Avg.	0.960	0.697	0.956	0.960	0.950	0.463	0.940	0.972

=== Confusion Matrix ===

a	b	<-- classified as
80777	198	a = Attack
3212	1159	b = Normal

Time taken to test model on supplied test set: 2.87 seconds

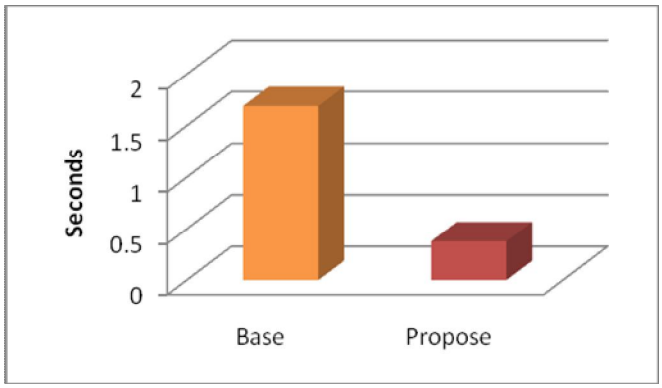

Fig. 3 Time taken to build model

In our analysis we used A2DEUpdatable and RepTree algorithm to build ensemble classifier. We build the ensemble classifier using WEKA machine learning tool. 10 fold cross validation is used for testing the classifier.10 Trees are used to build the RepTree. Following evaluation indices are defined from the confusion matrix. The confusion matrix shows the distribution of instances that are either attack or normal.

1) Accuracy= TP+TN/ (TP+FP+FN+TN)
2) False alarm rate =FP/ (FP+TN)
3) Detection Rate= It is the ratio between total numbers of attacks detected by the system to the total number of attacks present in the dataset DR= TP/TP+FN

VII. CONCLUSION & FUTURE SCOPE

In this research paper, we proposed a novel ensemble classifier (RepTreeA2DEU) for intrusion detection system.

The proposed approach efficiently classifies network traffic as normal or malicious. The results indicate that proposed classifier is accurate than RF and AODE classifiers. We considered Kyoto data set for experimental analysis. As Base classifiers are not capable of detecting the attacks accurately, proposed Ensemble classifier outperforms base classifiers A2DEUpdatable and RepTree. The results presented in this paper show that integration of A2DEUpdatable, RepTree and pre-processing technique will yield the good result for IDS. In the future, different datasets and parameters are taken into consideration for understanding the concept of IDS more efficiently and compare them to show illustrate the effective of the technique.

REFERENCES

[1] Vinutha H.P , Dr.Poornima B, "A Survey - Comparative Study on Intrusion Detection System" , International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 7, July 2015.

[2] K Ishthaq Ahamed, Dr. Shaheda Akthar, "Survey on Artificial Neural Network Learning Technique Algorithms" © 2016, IRJET.

[3] A. Shrivastava and S. Hardikar (2012, July). Performance Evaluation of BPNN and Genetic Algorithm. VSRD International Journal of CS & IT [Paper]. 2(7), pp. 621-628.

[4] V. Jaiganesh , S. Mangayarkarasi , Dr. P. Sumathi , "Intrusion Detection Systems: A Survey and Analysis of Classification Techniques" , International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 4, April 2013

[5] Okta Nurika , Nordin Zakaria , Fadzil Hassan , Low Tan Jung, "Workability Review of Genetic Algorithm Approach in Networks", 978-1-4799-0059-6/13/$31.00 ©2014 IEEE

[6] Ms. Snehal S. Joshi , Mr. Navnath D. Kale, "Survey: Support Vector Machine and Its Deviations in Classification Techniques" , © 2014, IJARCSSE.

[7] Dr. Shashikant Dugad , Vijayalakshmi Puliyadi , Heet Palod , Nidhi Johnson , Simran Rajput , Swapna Johnny, "Ship Intrusion Detection Security System Using Image Processing & SVM" , 978-1-5090-2794-1/17/$31.00 ©2017 IEEE.

[8] Om Prakash Nirankari, Prakash Pawar, Kotaro Kataoka, "Optimizing Service Chain ID Generation for Flow Rule Compression" , 978-1-5090-0933-6/16/$31.00 ©2016 IEEE.

[9] Anand Keshri , Sukhpal Singh , Mayank Agarwal ,and Sunit Kumar Nandi, "DoS Attacks Prevention Using IDS and Data Mining" , 978-1-5090-4291-3/16/$31.00c2016.IEEE.

[10] HosseinGharaee,HamidHosseinvand, "A New Feature Selection IDS based onGeneticAlgorithmandSVM" , 978-1-5090-3435-2/16/$31.00 ©2016 IEEE.

[11] GözdeKarataş, "Genetik Algoritma ile Saldırı Tespit Sistemi Genetic Algorithm For Intrusion Detection System" , 978-1-5090-1679-2/16/$31.00©2016IEEE.

[12] Mr. Kamlesh Patel, Mr. Prabhakar Sharma, "An Implementation of Intrusion Detection System Based on Genetic Algorithm", ISSN (Online) 2278-1021/ International Journal of Advanced Research in Computer and Communication Engineering ISO 3297:2007