ENHANCED RFA : A METHOD FOR EFFICIENT INTRUSION DETECTION

[1]Mr. Vishnu Muraleedharan, [2]Mr. Subin Omanakuttan

[1] *PG Scholar* , [2]*Assistant Professor*
[1]Department of Computer Science,
[1]College Of Applied Science (IHRD), Mavelikara, Kerala, India

Abstract: The usage of internet has been increasing day by day .Internet influences our privacy a lot. We share our important data to an intended person or to someone we don't know. We believe that our shared data will securely reach to the recipient. But it's only our expectation. There may be a chance that some of our shared data may be corrupted. The proposed system is designed to prevent intrusion and shares data securely between client and server using RFA and SVM algorithm. The random forest algorithm is designed to classify and compare other system over the networked data. And SVM's are used for classification and regression. The information gain method was used to improve the accuracy of RFA. To calculate the performance NSL KDD data set has been used. The NSL KDD is a data set available in many clustering algorithms. It is a data mining tool. The result of proposed system is compared with the other existing system. And we will get the result that the proposed system is far better than the previous systems.

IndexTerms **- Privacy,RFA, SVM, NSL KDD**

I. INTRODUCTION

Establishing security over internet is really a concern because of the rapid usage of internet. An intrusion detection system is used to detect the malicious activities. The network intrusion detection system is developed at a strategic point or points with in the network, where it can monitor inbound and outbound traffic to and from all the devices on the network. Historically intrusion detection systems were categorized as passive and active. Passive IDS that detect malicious activity and would generate alert or log entries, but wouldn't take an action. An active IDS, sometimes called an intrusion detection and prevention system. It generates alerts and log entries but also be take actions. An intrusion detection system may implement as a software application or as a network security. Various types of intrusion systems are:

- Network intrusion detection system (NIDS)
- Host based intrusion detection System (HIDS)
- Perimeter intrusion detection system (PIDS)
- VM based intrusion detection system (VMIDS)

The proposed system is introduced with Big Data, is a large set of information, and it can be categorized as unstructured or structured. Structured data consist of information already managed by the organization in database and spread sheets. It is numeric in nature. Unstructured data is information that is unorganized and does fall into a predetermined model. Big Data is a data with huge size and is growing, IDS is used to solve the problem in the network security and it is introduced in Big Data. That is network data don't have a form that's why Big Data is introduced here.
This system is introduced as a method for an intrusion detection scheme which is used on a network over random forest and SVM algorithm. It is used to detect the intrusion with more accuracy.

II. RELATED WORK

At present a number of researches are going on the field of effective intrusion detection. With the advance of technology, intrusion detection is become a herculean task. Some important papers that seriously discussed this issue is included in this section

- Feature selection based hybrid Ids using K means and radio basis function. This system is proposed by ujwala Ravle. It is a hybrid technique which combines both K means and SVM. KDD cup99 data set is used for experimental results.
- A hybrid intelligent approach for IDS, it is proposed by Mrutynjaya pande et al. In this system a combination of classifiers are used to improve the resultant model. It use a classification strategy with tenfold cross validation. The results are show in the base of NSL KDD data set.
- Multistage filtering for network IDS, it is proposed by P Natesan et al. In this system an enhanced ad boost with decision tree algorithm and a naïve Bayes used to detect frequent attack.
- IDS using random forest and SVM, it is proposed by Md AL Mehidi Hassan et al. It uses two models for IDS using SVM and Random forest. The performance of these methods are compared based on their accuracy, false negative precision

Network IDS using Random forest and PSO. It is proposed by Arif Jamal et al in 2012. Here use binary PSO which is used to select appropriate feature for classifying intrusion. In this system Random forest algorithm. It has two stages feature selection and classification. This method implemented in MATLAB..

III. PROPOSED WORK

Intrusion detection system is a networked system used to monitor network traffic and alert it. Once an attack is monitored, it shows an abnormal behaviour, and this information is passed to the administrator. The administrator will do further work on that

data. The existing system doesn't properly checks the security of the packet, accuracy of the packet, and dimensionality of the data set. These are the main issues faced by existing system on Big Data. But in the proposed system, it solves almost all issues in Big Data by implementing machine learning algorithms. Here we introduce two machine learning algorithms, Random Forest and SVM. This system check the dimensionality of data, and it provide the better results. But in the case of existing system they use only one algorithm is used to check the intrusion and analyse data. In the existing system, the client only sends the packet to the server and it may check the accuracy of the packet to find the intrusion. But it is not an efficient way to check the accuracy and find the intrusion. It cannot compare the accuracy and with other algorithms. The proposed system is work based on two algorithm Random forest and SVM. The advantages of the proposed system are:

❖ **Checks The Accuracy**

The accuracy of a data is related to the quality of data. If the data has quality it must satisfies the requirement of its intended user. To satisfy the requirement of its intended user, the data must be accurate. The data accuracy refers to whether the data values stored in an object that must be correct. The data value to be correct, it must right value and can be represented in consistent form and it must be unambiguous form. The proposed system is accurate so it has quality. To check the accuracy some process are implemented they are: pre-processing, feature selection and it also use some classifiers for this purpose.

❖ **Finds The Intrusion**

Intrusion detection is the important feature of the proposed system. An IDS is work on the basis of accuracy. IDS can be introduced to find external operational faults. And it is defined in terms of Maliciousness. It can be used to find various types of attack. And can be used to detect some security issues. The IDS is considered as classification problem. The main aim of IDs is to detect intrusion, when sending packets through the network from client to server. And this can be done by using pre-processing and feature selection.

❖ **Provides Encryption**

The existing system will not provide any security to the data. The data set is passed to the server from client and it is not providing any encryption or security to that data. It only detecting the accuracy of that data. And make an assumption that the client is an intruder or not. But in this system we provide an additional security that the encryption to the data. That is, if a client will send data to the server and accepts those data. That is, if the client is not an intruder it will accept those data. And if it is an intruder the data will neglect. If the client is not an intruder the data is stored as an encrypted format in the server. By doing this an attacker cannot take the data from the server. Thus by providing an extra level security. If an attacker is tried to attack the server, the accepted data in the server is an encrypted form. So it will be very difficult for the attacker to steal the data from server.

IV. METHODOLOGY

The intrusion detection over Big Data is a real concern. The experiment is based on the data sets. The data sets are networked data, which provides different form in different time. Sometimes it will normal and some time it will abnormal, that is the networked data show normal and abnormal behaviour. It is solved by using pre-processing stages and classified using random forest and SVM algorithms. And an Information Gain method was used to evaluate the accuracy of data. Finally the designed data is transmitted from client to server using both these algorithm. And there will be a comparison of algorithms. After the comparison, we will get clear idea that the proposed system is much better than different existing classifiers.
The detailed description of each step is mentioned below.

a. Pre processing

The pre-processing is the initial stage of our proposed system. The data set, that we have is converted or controlled into a specific format. So here we use networked data that wouldn't have a specific format and transforming raw data into an understandable format. The raw data are real world data; they are inconsistent, incomplete in nature. So Data pre-processing is used to solve these issues. The pre-processed data is further taken into another processing. In our system pre-processing is taken on Big Data and is done in the intrusion detection system. So the pre-processing stages are really needed in Big Data in intrusion detection system. It is really important for the machine learning algorithms for the classification. In this stage in order to get the specific feature of our data, Information Gain method was used.

Information Gain is a type of pre-processing procedure. Information Gain(IG) method measures how much "information" a feature gives as about the class. Information Gain is the main key that is used decision tree algorithms in order to generate the decision tree. The Information Gain is actually measuring the entropy with respect to the class. The entropy is the measure of impurity, disorder or uncertainty in a bunch of examples. Information Gain method is the expected reduction in entropy caused by partitioning the examples according to a given attribute.
Attribute Selection Method: Information Gain

- Condition for stopping partitioning
  All tuples for a given node belong to the same class
- Attribute list is empty
  Majority voting is employed for classifying the deal.
- There are no tuples for a given branch Dj,

A leaf created with the majority class in D Select the attribute with the highest Information Gain

- This attribute minimizes the information need to classify the tuples in resulting partitions.
- Let Pi be the probability that an arbitrary tuples in D belongs to class Ci, estimated by |CiD|,|D|
- Expected information(entropy) needed to classify a tuple in D

InFo(D)= $-\sum_{i=0}^{m} \text{Pi} \log_2(Pi)$

- Entropy represents the average amount of information needed to identify the class label of a tuple in D.
- Attribute A has V distinct

A can be used to split D into V partitions, where Dj contains those tuples in D that have outcome aj of A.
If A is selected, we wish each partitions Dj is pure.

- Information needed (after using A to split D into V partitions) to classify D.

$$\text{Info}_A(D)=\sum_{j=1}^{v} \frac{|Dj|}{|D|} * info(Dj)$$

- The smaller the information needed the greater the purity of the partitions.
- Information Gained by branching on attribute A

$$\text{Gain}(A)=\text{Info}(D)-\text{InFo}_A(D)$$

b. Feature Selection

Feature selection is a Data Mining tool. It is commonly used in pre-processing and is a part of machine learning. Feature selection is also called attribute selection. It is a procedure of selecting attribute in our data such as tabular data. It is done by dividing the data into subsets and analyse each of these subset make the decision. That is, it selects subset that has most relevant data. Feature selection reduces the attribute data set and sometimes includes and exclude without changing them.

V. About The Algorithm

In our system, we are implementing two algorithms both are classifiers ,one is Random forest and the other one is SVM. Here we are actually doing the comparison of two algorithms. Based on the algorithm we detect the accuracy of data. And sends it to the server using both the algorithms. Thus by doing, we get clear idea about which algorithm is doing better in Intrusion Detection System.

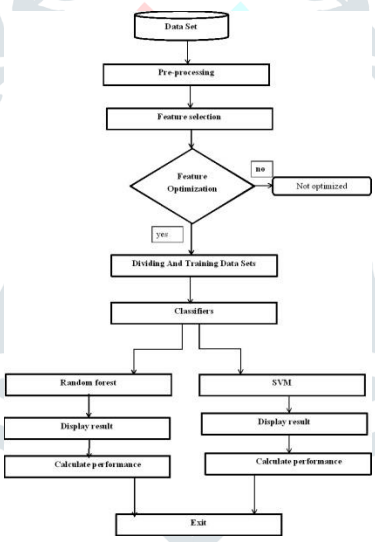

Fig 1: Proposed model

a. Random Forest Algorithm

Random forest is one of the most popular classification algorithm. In machine learning it is normal us random forest classifier. Random Forest algorithm can be used both for classification and regression problems. Random Forest algorithm is a classification algorithm, as the name suggests, this algorithm creates the forest with a number of tree. That is, consider a forest, if the forest has huge amount trees, then it look like a dense forest. Similarly in this, more number of trees will give high accuracy results. It is similar to the decision tree algorithm which builds several trees and calculates each of these outputs. Then joins these outputs to make general decision. Random Forest works by applying a bootstrap model to each of the tree from the original data set. The main advantage of Random Forest is that it has low error rate in classification

.

❖ **Advantage Of Random Forest Algorithm**

- Random Forest can be used in both classification and regression problem.
- Random Forest is very helpful for finding missing values.

- It also useful in categorize the variables.
- Random Forest has less error rate.
- Random Forest is very easy.
- Random Forest is a flexible, easy to use machine learning algorithm.

❖ **Input To The Algorithm– NSL KDD Data Set**

The NSL-KDD Data Set is used to evaluate the IDS. The KDD cup99 data set is the base model. And the NSL-KDD is an updated version. Because of the base model has some limitations, NSL-KDD data set proposed by Tavallaee et al. The NSL-KDD has some difference with KDD 99 that is, it doesn't have the record of trained data set, and it doesn't maintaining duplicate record. The NSL-KDD data set are independent to the operating system; it collected data as networked data. At the NSL-KDD contain many types

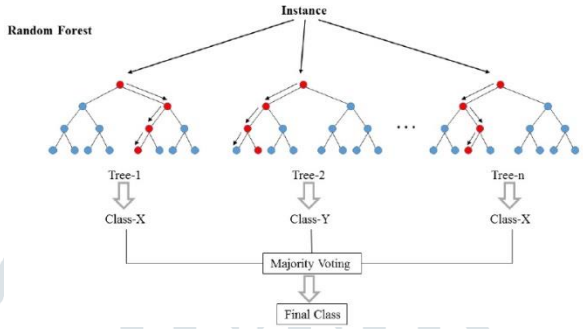

Fig 2: Random Forest Algorithm

b. Support vector machine

Support vector machine (SVM) is a supervised machine learning algorithm which can be used for both classification and regression. It is mostly used in classification problem. Each data item is plot as a point in n dimensional space with value of each feature will be the value of particular coordinate support vector machine algorithm is to find a hyper plane in N Dimensional space. It is used on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. To construct a hyper plane, SVM employs an interactive training algorithm which is used to minimize an error function: SVM models can be classified into 4 distinct group.

- Classification SVM Type 1

For this type of SVM, training involves the minimization of the error function.

$$\frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i$$

Subject to the constraints:

$$y_i \left(w^T \phi(x_i) + b \right) \geq 1 - \xi_i \text{ and } \xi_i \geq 0, i = 1, \ldots, N$$

- Classification SVM Type 2

In contrast to Classification SVM Type 1, the Classification Type 2 model minimizes the error function.

$$\frac{1}{2} w^T w - v\rho + \frac{1}{N} \sum_{i=1}^{N} \xi_i$$

Subject to the constraints:

$$y_i \left(w^T \phi(x_i) + b \right) \geq \rho - \xi_i, \xi_i \geq 0, i = 1, \ldots, N \text{ and } \rho \geq 0$$

Regression SVM

Y=f(x)+noise
The task is to find a functional form for that can be predicting new cases that the SVM has not been presented with before. This can be achieved by training the SVM model and a sample set.

- Regression SVM Type 1

For this type of SVM the error function is:

$$\frac{1}{2}w^T w + C\sum_{i=1}^{N}\xi_i + C\sum_{i=1}^{N}\xi^*_i$$

Which we minimize subject to:

$$w^T\phi(x_i) + b - y_i \le \varepsilon + \xi_i^*$$

$$y_i - w^T\phi(x_i) - b_i \le \varepsilon + \xi_i$$

$$\xi_i, \xi_i^* \ge 0, i = 1,\ldots,N$$

- Regression SVM Type 2

For this SVM model the error function is given by:

$$\frac{1}{2}w^T w - C\left(v\varepsilon + \frac{1}{N}\sum_{i=1}^{N}\left(\xi_i + \xi_i^*\right)\right)$$

Which minimizes subject to?

$$\left(w^T\phi(x_i) + b\right) - y_i \le \varepsilon + \xi_i$$

$$y_i - \left(w^T\phi(x_i) + b_i\right) \le \varepsilon + \xi^*_i$$

$$\xi_i, \xi^*_i \ge 0, i = 1,\ldots,N, \varepsilon \ge 0$$

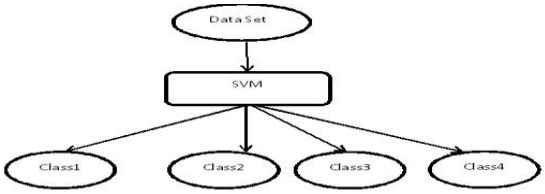

Fig 3: SVM Model

- Application

1. SVMs are useful in text and hypertext categorization.
2. SVM are also useful in image classification.
3. SVM can be used to detect hand written characters.
4. SVM are widely used in biological and other sciences.

Algorithms	FPM	TPM	Accuracy	Precision
SVM	0.006	0.957	94.97	0.987
KNN	0.009	0.933	93.57	0.975
REP Tree	0.005	0.988	98.1	0.721
Naïve Bayes	0.006	0.949	95	0.949
Proposed Model	0.001	0.993	99.5	0.993

Table 1: Performance Result of Proposed model

vii. Performance Measure

Performance measure is used to analyse the accuracy, FPM, TPM, Precision of the proposed model

True Positive Measure (TPM) = $\frac{TP}{\Sigma(TP,FN)}$%100

False positive Measure (FPM) = $\frac{FP}{\Sigma(TN,FP)}$%100

Accuracy = $\Sigma((TP,TN) \div (TP,FP,TN,FN))$

Precision=$\frac{TP}{\Sigma(TP,FP)}$%100

TN=True Negative
TP=True Positive
FP=False Positive
FN=False Negative

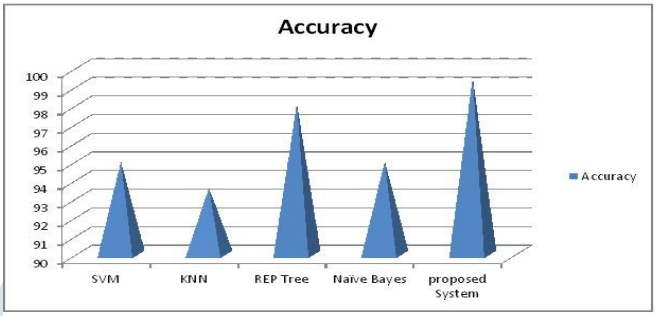

Fig 4: Performance Accuracy of Proposed Model

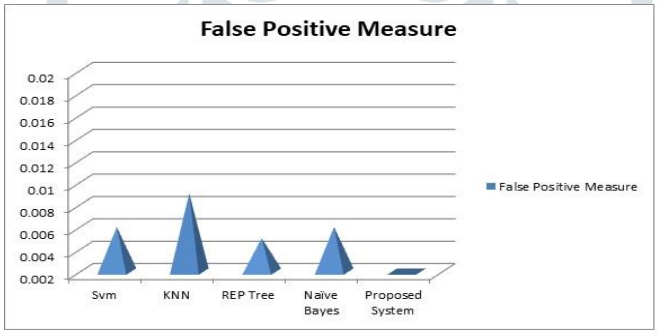

Fig 5: False Positive Measure of Proposed Model

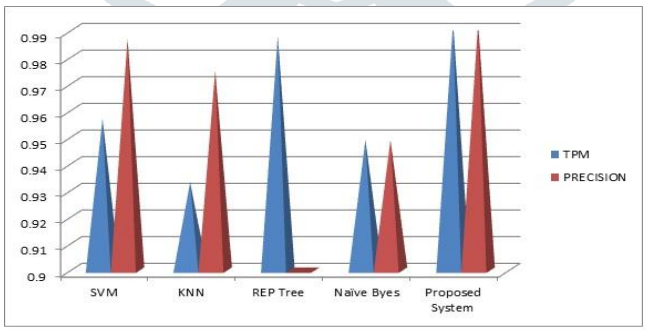

Fig 6: Performance TPM and Precision of Proposed Model

VI. Performance Analysis of Proposed Model

The comparison of proposed model is based on the accuracy of data over intrusion detection system. The main difference of our system with the existing system is that, the existing system is only implementing one algorithm, but in this system, two algorithms are implemented. One is Random Forest and other one is SVM. Both of these algorithms are classifiers. This system evaluates, FPMand TPM, which are the performance measures that are used to evaluate the data set with algorithms. The existing system does not provides any security to the data that client sends to the server. But in this system we provides an encryption to the data that resides in the server. The server maintains a list of client and their data's, which are in encrypted form. That is if there is any chance that an intruder will attack the server then It will very difficult for him to steal data. And also it provides summarized result which is better than all existing systems . It shows that proposed model is more accurate and takes less time to build and also from this it is obvious that error rate of our system is very less and the precision rate is very high

VII. Conclusion And Future Work

The main focus of this system is to improve the accuracy of data over intrusion detection using various algorithms. The data set which is chosen for our system is reformed through various stages, such as pre-processing and feature selection. The pre-processing stages includes Information Gain method, which evaluates and analyses the data set. This system is implemented in Big Data. And the pre-processed data is optimized and classified with random forest and SVM algorithm. The results of the algorithm are compared with one another. And finally obtained the accuracy of the data and generated a graph. On comparing with the existing systems, it is very efficient and the main advantage is that it provides encryption to the data that resides in the server. The proposed system is very user friendly and provides less error rate. It is observed that proposed model has outperformed the previous mechanisms devised for the same purposes.

VIII. Acknowledgment

We are thankful to our guide Prof. Subin Omanakuttan Assistant professor in department of computer science for his valuable support. Also we are thankful to our department for the technical support.

References

[1] R. Chitrakar, and C. Huang, "Anomaly based Intrusion Detection using Hybrid Learning Approach of combining k-Medoids Clustering and Naïve Bayes Classification," In Wireless Communications, Networking and Mobile Computing (WiCOM), 8th International Conference on, pp. 1-5, IEEE, 2012.

[2] G. V. Nadiammai, S. Krishnaveni and M. Hemalatha, A Comprehensive Analysis and Study in IDS Using Data Mining Techniques, *IJCA*,
vol. 35, pp. 51–56, November–December (2011).

[3] L. Breiman, Random Forests, Machine Learning, vol. 45, no. 1, pp. 5–32, (2001).

[3]. Dokas, P., Ertoz, L., Lazarevic, A., Srivastava, J., & Tan, P. N., "Data Mining for network intrusion detection", Proceeding of NGDM, pp.21–30, 2002.

[4]. Wu, S., and Yen, E., " *Data mining-based intrusion detectors*", Expert Systems with Applications, vol.36,no.3, pp.5605–5612.,2009.

[6] Mary Slocum "Decision making using ID3" RivierAcadmic Journal, Vol 8, No 2, 2012.

[7] Dewan Md. Farid, Jerome Darmont and Mohammad Zahidur Rahman" Attribute Weighting with Adaptive NBTree for Reducing False Positives in Intrusion Detection" International Journal of Computer Science and Information Security, Vol. 8, No. 1, 2010 PP 19-26.

[8] Santosh Kumar Sahu Sauravranjan Sarangi Sanjaya Kumar Jena, "
A Detail Analysis on Intrusion Detection Datasets", 2014 IEEE
International Advance Computing Conference (IACC)

[9] Sapna S. Kaushik, Dr. Prof.P.R.Deshmukh," Detection of Attacks in an Intrusion Detection System", International Journal of Computer Science and Information Technologies, Vol. 2 (3), 2011,982-986

[10] Kezunovic M, Xie L, Grijalva S (2013) The role of big data in improving power system operation and protection. In: Bulk Power System Dynamics and Control - IX Optimization, Security and Control of the Emerging Power Grid (IREP), 2013 IREP Symposium. IEEE, Rethymno, Greece. pp 1–9

[11] Denning DE (1987) An intrusion-detection model. Software Engineering IEEE Trans SE-13(2):222–232. doi:10.1109/TSE.1987.232894

[12] M. A. Jabbar and B. L. Deekshatulu, Priti Chandra, Alternating Decision Tree for Early Diagnosis of Heart Disease, *IEEE*, pp. 322–328,
(2014).

[13] Jehad Ali, *et al.*, Random Forest and Decision Trees, *IJCSI*, vol. 9, no. 3, pp. 272–278, (2012).

[14] Lidong Wang*, Randy Jones "Big Data Analytics for Network Intrusion Detection: A Survey. International Journal of Networks and Communications 2017, 7(1): 24-31 DOI: 10.5923/j.ijnc.20170701.03

[15]. http://arxiv.org/abs/1201.1587.

[16] Rachana Sharma & Priyanka Sharma, Preeti Mishra & Emmanuel S. Pilli "Towards MapReduce Based Classification approaches for Intrusion Detection". Intrnation conference 2016 IEEE PP-361-366

[17] Miss Gurpreet Kaur Jangla1, Mrs Deepa.A.Amne2.." Development of an Intrusion Detection System based on Big Data for Detecting Unknown Attacks. International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 12, December 2015