MAIDEn: A Machine Learning Approach for Intrusion Detection using Ensemble Technique

Habil Damania
Dept. of Computer Engg.
M.E.S College of Engineering
Pune, India

Aditya Jagtap
Dept. of Computer Engg.
M.E.S College of Engineering
Pune, India

Abhishek Jain
Dept. of Computer Engg.
M.E.S College of Engineering
Pune, India

Chaitanya Chavan
Dept. of Computer Engg.
M.E.S College of Engineering
Pune, India

Shraddha Khonde
Assistant Professor
M.E.S College of Engineering
Pune, India

ABSTRACT

An Intrusion detection system is a machine or software that monitors the traffic in a network and on detection of a malicious packet, informs the user or a specific acting unit which can take further action and avoid the malicious packet from entering the network. This paper discusses a way to implement an intelligent IDS which classifies the normal traffic in a network with abnormal or attacked ones. This paper explains the method used to generate such a system and the various classifiers used in the generation process.

The proposed system of Intrusion Detection, classifies data with three different classifiers and an Ensemble technique which selects the majority of the three classifiers to assign the packet in the network as anomaly or normal. The dataset used to train the classifiers is the NSL – KDD dataset. The IDS proposed serves many applications in the field of Military Systems, Banks and Social Networking websites where data is very sensitive. The paper also explains related work done in this field and briefly explains every classifier, the network attacks and the dataset.

General Terms

Network Security, Intrusion Detection, IDS, Artificial Intelligence, Machine Learning, Ensemble, SVM. Random Forest

Keywords

IDS, Intrusion Detection System, Artificial Intelligence, AI, Majority Voting, Ensemble Learning, Random Forest, SVM

1. INTRODUCTION

Intrusion has become a growing concern today. With the advent of new technologies each day and widespread of computers (from personal computers to embedded systems), security has become a very important issue. To name a few Attacks like Ransomware, Ddos, U2R, R2L have become a great deal of concern to every computer in the network. Such attacks compromise the security of the computer and obtain access to sensitive data. Hence, Security of any network is a high priority issue which must be taken care of. Various Intrusion Detection Systems (IDS) exist which help identify threats in the system but only an intelligent system will correctly yield them with maximum accuracy.

With Artificial Intelligence becoming pervasive in the computer world, it sets its foot into the area of Network Security as well. Hence, one could make full use of it and create a system that could provide a secure environment for the users in a network.

The aim is to create such a system. The above mentioned issues motivates all to select this project. The Aim is to create a Novel IDS which incorporates the methodologies of Machine Learning to identify the attacks in the network correctly with very less number of misclassifications which otherwise would go unidentified in traditional Intrusion Detection Systems. The following sections will discuss the scope of the project and the technologies that will be used to achieve the final product.

2. PROJECT SCOPE

The aim is to make a system which will classify each packet in the network as anomaly or attack. On Detection of an attack the user or network administrator will be alarmed either through a computer software or an Android application so that appropriate actions can be taken.

Before starting the implementation an appropriate dataset must be selected to train the classifiers. The Dataset selected is the NSL – KDD. The next step is to select the classifiers. Random forest, Support Vector Machine and Convolutional Neural Networks are the three classifiers that are selected. After selection of classifiers, modeling and training is performed. For obtaining better accuracy Ensemble Technique is used, which performs majority voting amongst all the classifiers to select the class of the data. The use of majority voting makes the system stand out amongst other Intrusion detection Systems. Since, three classifiers are used in conjunction, a misclassification by one classifier will be correctly classified by the other two. It will improve the accuracy of detection significantly.

3. LITERATURE REVIEW

Preeti Aggarwal and Sudhir Kumar Sharma [1] gave us in depth knowledge of the KDD-99 dataset .They separate the attack types into 4 types Basic, Content, Traffic and Host. The 2 main evaluation metrics ,Detection Rate and False alarm rate .After clustering into 4 types using all attributes, 15 subsets were created. Each class dominance was then used to improve Detection rate and False alarm rate. The main aim was to find higher DR and decrease false alarm rate.

Jayshree Jha and Leena Regha [2] created an SVM classifier and the feature selection in the classifier was done by using a combination K means and Information gain. Information gain gives us the importance of each feature .It is discovered that the difference between choosing top 23 features and top 30 makes a difference of difference of 0.05%.They first rank features of based on information gain and then select features using K-means algorithm.

Nabila Farnaz and M.A Jabbar [3] modeled a random forest classifier and compared it with j48 classifier. The modeled the classifier using NSL-KDD dataset. The dataset was first clustered by using the classes of attacks after pre-processing. In preprocessing, for reducing features they used feature selection by finding out symmetrical uncertainty measure. In classifier training they created 100 trees .After classifier training, they compared the results using detection rate and false alarm rate. The accuracy was 99.67%, Detection rate was 99.83 and false alarm rate 0.00527%.The used 10 cross validation method.

Md. Al Mehdi Hasan, Md. Nasser, Biprodip Pal and Shamim Ahmad [4] created IDS using SVM and Random Forest classifier .After comparing results they found out Random Forest is better in term of computational time as it is faster than SVM and it produces similar accuracy to SVM .They used radial basis kernel in this SVM and the accuracy for testing dataset is 92.99 and the Random Forest accuracy for testing in 91.41%.The precision for random forest was 10% more than even while the process time was less.

Hee-su Chae, Byung-oh Jo, Sang-Hyun Choi, Twae-kyung Park [5] proved that in KDD dataset all the 41 features are not relevant by using, information gain, Gain ratio and Correlation based feature selection. The classifier used was decision tree classifier. They proposed a method AR (Attributed Ratio) which is a new method for giving importance of classes and compared with GR, IG and CFS. They proved that by using only 22 features out of 41, an accuracy of 99.79 can be achieved.

4. DATASET USED

4.1.1 KDD

KDD 99 dataset is the most reliable dataset for network security. Most of datasets used for network security are derived from this dataset such as NSL-KDD, Corrected-KDD, 10% KDD etc. It has over 300 thousands of entries and all of them are labeled .Even though the data set was created in 1999, it is the most preferred dataset. There are 4 main species of attacks and around 35 subspecies and covering most of the network related attacks

5. CLASSIFIERS USED

5.1 Random Forest (RF)

Random forest is an ensemble classifier. It has a higher classification accuracy compared to single decision tree. Random forest contains many decision trees which are trained with the same dataset and different features selected at random. It avoids overfitting as features and data are randomly selected.

5.1.1 Training

The goal of Random Forest is to use distributed approach for classification. The features used to train decision trees will be selected using their importance in the KDD-99 dataset. The trained model will be placed on a distributed network to increase real time performance and reliability.

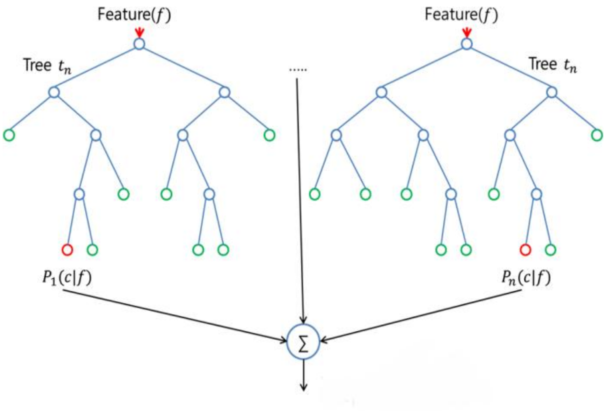

Fig 1 Feature Selection in Random Forest

The class assigned by the decision trees will be put to a vote and majority class will be assigned to the input data

$$(c|f) = \sum_{1}^{n} P_n\,(c|f)$$

The number of trees in the forest will depend on the current network traffic and processing capabilities available. Maximum idle processing power can be used to increase number of trees in the forest which will give more precise classification

5.2 Support Vector Machine

In SVM classifier training, creation of 'n' number of hyperplanes for 'n' number of classification is done.
SVM can be used for both, classification and regression but in this case it will be used for classification. Hyperplane is a plane which divides two different classes of nodes in a space and hence classification is done.

5.2.1 Steps for Generating SVM

5.2.1.1 Data Preprocessing

SVM's are incapable of processing categorical data since they only process numerical data.
In order to train SVM from KDD dataset conversion of string data into appropriate numerical data needs to be done for training the classifier. The process of conversation needs to be saved in order to test the live data because classifier will not work if live data is not converted according to the conversation process of training data.

Algorithm:
1) Scan string value.
2) Check if numerical value is assigned.
If assigned replace string with numerical value.
3) If not assigned assign value and replace string.
4) Save the replacing values for future use.

5.2.1.2 Data Normalization

Data normalization is extremely important for training classifier using KDD dataset as range of value for each feature varies a lot. If data is not normalized, then it may occur that the trained classifier would be biased to certain features only and also training time increases
And the accuracy decreases and also the value with which the data is divided must be saved in order divide the live data to classify it.

$$N_2 = (N_1 * min)/(max - min)$$
$$where,\ N_2 = New\ Value$$
$$N_1 = Old\ Value$$

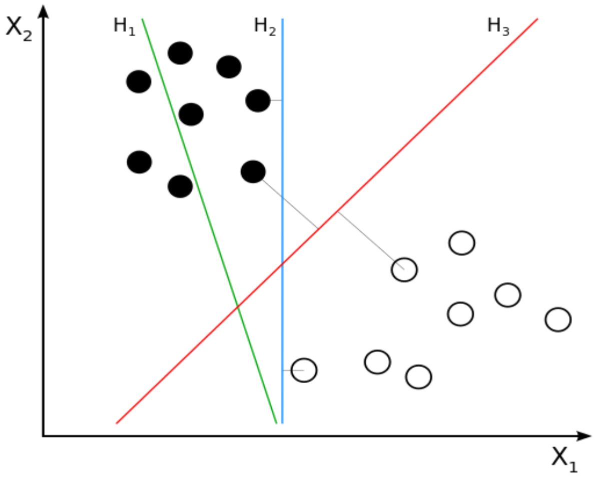

Fig 2: SVM represented by Hyperplanes H1, H2, H3

5.3 Ensemble Learning

Ensemble learning is a method to improve the accuracy of a model by creating a set of several Machine Learning Classifiers and then predicting the most fitting class of the input data. This is done by making use of a majority voting scheme between all classifiers and selecting the class of the data which has most votes that is the class which majority of the classifiers have predicted. Ensemble Techniques could be Bagging, Boosting or Bayes Optimal Classifier.

In this paper, Boosting is used. Boosting attaches weights to each data. One potential drawback of boosting is overfitting.

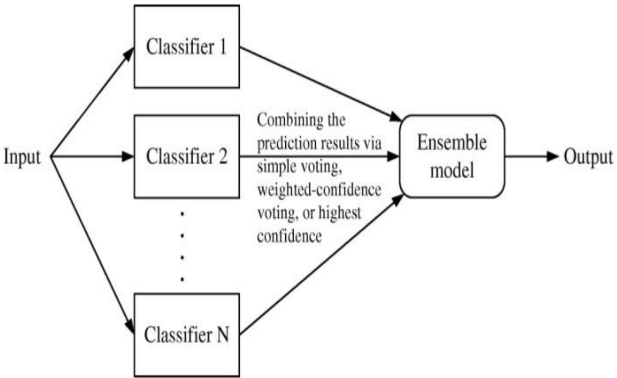
Classifier 1
Input
Classifier 2
Combining the prediction results via simple voting, weighted-confidence voting, or highest confidence
Ensemble model
Output
Classifier N

Fig 3: Transition of an Ensemble Method

6. CONCLUSION

In this paper, the aim was to discuss the various machine learning classifiers for Intrusion detection. An important technique called Ensemble Learning was discussed which if used decreases the misclassification rate significantly. The other related work done on similar grounds was discussed as well.

7. FUTURE SCOPE

Various measures can be taken to improve the system in the future. Better Machine Learning Algorithms which take less processing time can be used. Also Boosting can be replaced by a better Ensemble Technique or the overfitting which takes place in boosting can be reduced.

8. ACKNOWLEDGMENTS

Our sincere thanks to Prof. S.R. Khonde for helping us in the development of this paper and inculcating her knowledge in this domain to us. Her constant motivation and support at every stage of development has helped us design this paper.

9. REFERENCES

[1] Preeti Aggarwal, Sudhir Kumar Sharma ICRTC 2015. Analysis of KDD Dataset Attributes. .

[2] Jayshree Jha and Leena Ragha ICWAI 2013. Intrusion Detection System using Support Vector Machine.

[3] Nabila Farnaaz and M. A. Jabbar. IMCIP 2016. Random Forest Modeling for Intrusion Detection System Tavel, P. 2007 Modeling and Simulation Design. AK Peters Ltd.

[4] Md. Al Mehedi Hasan, Md Nasser, Biprodip Pal and Shamim Ahmad. JILSA 2014. Support Vector Machine and Random Forest Modeling for Intrusion Detection System Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. 3 (Mar. 2003), 1289-1305.

[5] Sang-Hyun Choi, Hee-su Chae, Byung-oh Jo and, Twae-kyung Park .Feature Selection for Intrusion Detection and NSL-KDD.