A Novel Approach to Solve Class Imbalance by using Ensemble Classifier

Dr. D. Sivakumar[1], Dr. S. M. Uma[2],
Kings College of Engineering, Punalkulam,
Thanjavur, Tamil Nadu

.

Abstract - Security is a key controversy to both computer and computer networks. An Intrusion Detection System is a software that superintend a single or a network of a computers for denmastry activities which are pursued at purloining or inspecting information or deprave network protocols. IDS can be grouped into Signature based Detection (SBD) and Anomaly based Detection (ABD). Machine Learning Techniques have been scrutinized and emulated in label of their detection potentiality for identifying the different groups of attacks. In this Paper, we Proposed a comprehensive evaluation of diverse machine learning techniques for locating the root of complications in recognizing Intrusion Activities. Controversies that are analogous to discerning low-frequency attacks utilizing network attack datasets are also explored and effective methods are recommended for betterment. Numerous Data Mining tools for Machine Learning have also been incorporated in this paper. By using Sampling Technique, the efficiency and scalability was improved better compared to formal approaches.

Keywords: Intrusion Detection System, Machine Learning, Precision, ROC, True Positive, False Negative

INTRODUCTION

A lay of skill used for perception of anomalous etiquette of networks. Based on the speculation that the etiquette of intruder is contradictory from that a usual user. As the elegant attack intensifies, the skillful Intrusion Detection approach is essential to overcome the annoying activities. In Common ,the potency of IDS is a survey of its proficiency to identify intrusion, to the least those that could possibly cause detrimental destruction. Few common parameters for estimates are detection rate, false positive, false negative, true positive ,false alarm. Much of the Existing strategy focused on upgrading the detection rate and therefore to some extent, the field has been massively well researched. In this Paper, we inspect an aggregate of ABD methodologies that has been developed for IDS. Each Method was tested using various available datasets targeting a number of attacks. Our main review is to find the key advantages of each technique as well as their drawbacks. In Succeeding period, this paper can benefit as a reference point and furnish scope to improve the existing approach for further research.

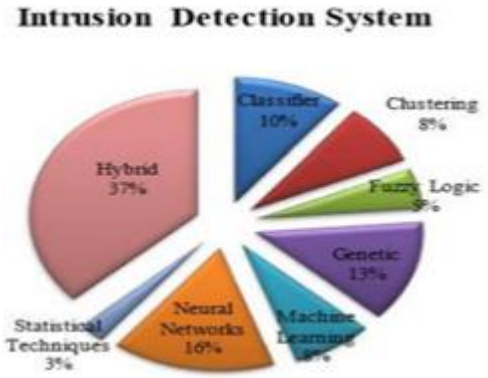


Figure 1.1

Machine Learning based IDS contributes a Learning based system to find category of attacks based on the learned normal and attack performance. The ultimate aim of machine learning based IDS is to imitate a common representation of known attack. Anomaly based IDS are depends on the speculation that attacker behavior differs from normal users' behavior which helps in identifying the enlarging attacks.

Single classifier

Single machine learning classifier can be used to address the problem of intrusion detection. Several Techniques such as Support Vector Machine (SVM), Self-Organizing Maps (SOM) and K-Nearest Neighbor (KNN) have been used to resolve the problem and the results have shown some significant achievements. The data sets are pre-processed to be used by SVM classifier. SVM is trained over the training dataset and as a result, decision model is generated.

Strategies in Machine Learning

a) Artificial Neural Network

Neural Network Learning methods impart a powerful approach for approximating real valued, discrete-valued and vector-valued target functions. Neural Networks are suitable for the problems where a) Instances are represented by many attribute-value pairs. b) Training sample may contain errors. c) The learned function is typically difficult to understand by humans and this ability to understand the learned target function is not important by human.

b) Fuzzy Logic

Fuzzy logic is a mould of many-valued logic that deals with approximate rather than fixed and exact reasoning. Fuzzy logic can interpret the properties of a neural network and a precise description of its performance can be obtained. Neuro-fuzzy is very popular in the area of Intrusion Detection.

c)Ensemble of Classifiers/ Ensemble Learning

Ensemble learning makes use of multiple learners and combines the predictions made by a set of classifiers called as base learners. The use of multiple machine learning algorithms helps in generating a set of hypotheses for a problem. Some of the ensemble methods make use of the homogeneous base learners in which multiple instances of the same machine learning algorithm are used to generate a set of hypotheses over different sub-samples of the same training dataset.

Related Works

a) Intrusion Detection

Intrusion Detection Systems are applications that monitor a certain disk or program or network activities for malicious activities or violations of IT policy and produces reports to a management station or administrator if they match a certain signature. [1]. intrusions constitute only a small percentage of the total network and computer usage. The data that come in a streaming fashion requires online analysis. [2]. a misuse detection model is built based on the C4.5 decision tree algorithm and then the normal training data is decomposed into smaller subsets using the model. [3]

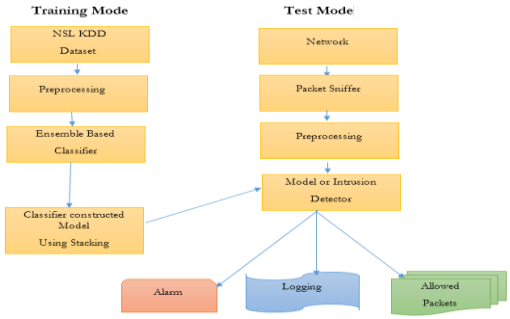
b) Network Security

Network forensics is scientifically proven techniques to collect, detect, ident, examine, correlate, analyze and document digital evidence from multiple sources to identify suspicious entities and stepwise action [4]. Monitoring of network traffic is an essential activity for network defenders in order to observe, analyze and finally identify any anomalies occurring in the network Rapid incidences of malicious attempts to compromise the confidentiality, integrity and access control mechanisms of a system or to prevent legitimate users of a service from accessing the requested resources have led to an increased demand for developing useful tools to visualize network traffic in a meaningful manner to support subsequent analysis.[5] .The network is also a pathway for intrusion.[6].

c)Algorithms

Genetic Algorithm is powerful because of some of its nice properties, e.g., robust to noise, no gradient information is required to find the global optimal or sub-optimal solution, self learning capabilities etc.[7] .K-means Clustering Algorithm is useful in describing the cluster formation in terms of attributes contribution to different cluster that is tested on various synthetic and real datasets to show its effectiveness.[8] The K-means algorithm, starting with k arbitrary cluster centers in space, partitions the set of giving objects into k subsets based on a distance metric. The centers of clusters are iteratively updated based on the optimization of an objective function. This method is one of the most popular clustering techniques [9]

SYSTEM ARCHITECTURE:



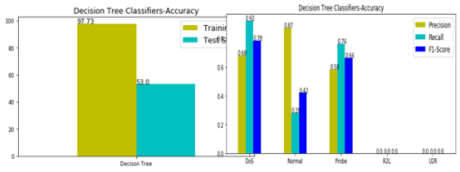
System module

Machine learning approach work in two phases: training set and testing set. In training phases, they perform the mathematical calculations over the training dataset and learn the behavior of traffic over a period. In the testing phases test occurrence is divided as normal or intrusive based on the well etiquette. Different favored machine approaches are chronicled below:

Module 1

Decision Tree

Decision tree is a type of supervised learning algorithm (having a pre-defined target variable) that is chiefly used in classification problems. Its activity for both categorical and continuous input and output variables. In this technique, we split the population or sample into two or more homogeneous sets (or sub-populations) based on most significant splitter / differentiator in input variables.



Module 2

K Nearest Neighbor

K Nearest Neighbours based classification is a type of lazy learning as it does not attempt to construct a general internal model, but simply stores instances of the training data. Classification is computed from a simple majority vote of the k nearest neighbours of each point, as shown in figure 1.2

Advantages: This algorithm is simple to implement, robust to noisy training data, and effective if training data is large.

Disadvantages: Need to determine the value of K and the computation cost is high as it needs to computer the distance of each instance to all the training samples.

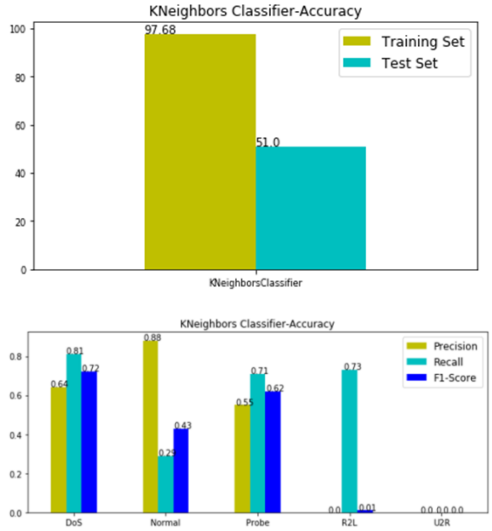


Figure 1.2

Module 3

Ensemble model

Ensemble learning is a model which is used to use to combines more than one weak learner are grouped iteratively to yield a better learner that can easily the training the samples from the given dataset. It can able to produce better performance when compare with standard classifiers.

Ensemble types

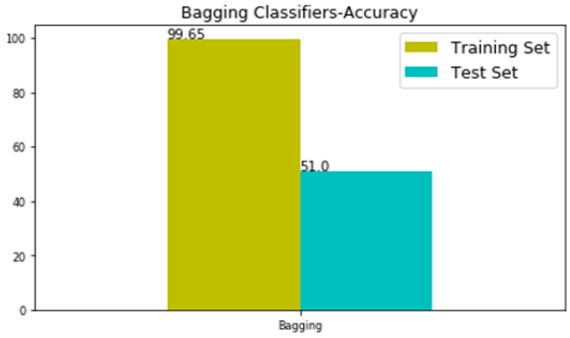
Bagging

Boosting –ADA boosting

Stacking

Bagging

Bagging tries to implement similar learners on small sample populations and then takes a mean of all the predictions. In generalized bagging, you can use different learners on different population. As you can expect this helps us to reduce the variance error. As shown in Figure 1.3



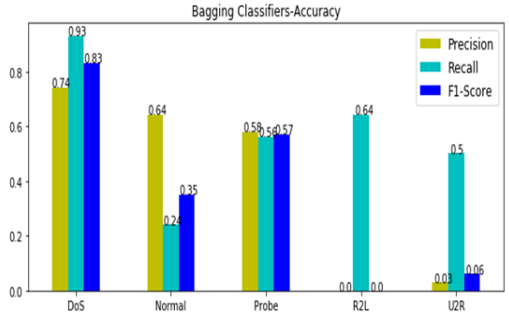


Figure 1.4

Boosting

ADA boost is one of the simplest and more reliable models in boosting. It basically works on the decision tree mechanism .It built the multiple decision trees models with sequential manner. Each tree correcting the errors from their previous tree model. In general, it uses weighted mechanism for giving high priority weight to incorrectly predicted instance and subsequent build the next level of tree to predict these values correctly. As shown in Figure 1.5

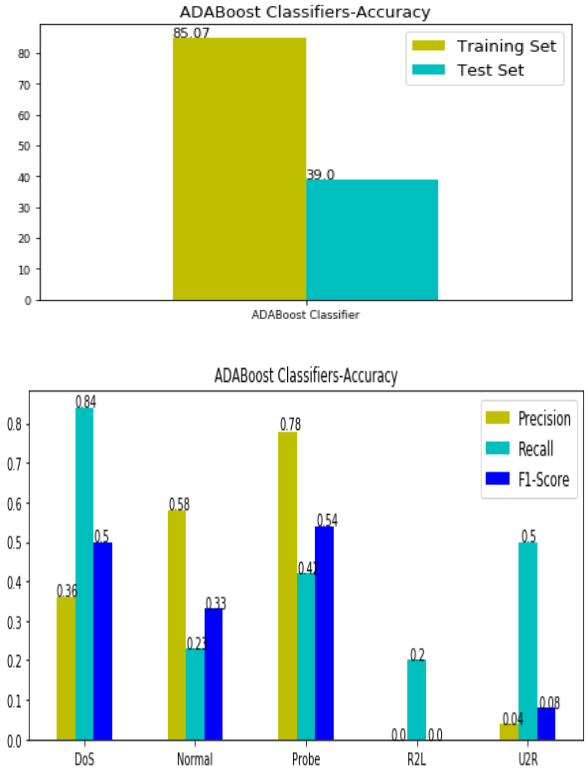


Figure 1.5

Stacking

Stacking is the ensemble approach it is also called (Meta Assembling) used to combine information from multiple predictive models to generate a new model. Stacking is

most effective when the base models are significantly different.

Stacking ALGORITHM:

Input: training Data D={xi, yi } ^mi=1;
Output: ensemble classifier H
Learn base level classifier
For t=1 to T do
Learn ht based on D
End for
Step 2:construct new dataset of predictions
For i=1 to m do
Dh= {xi, yj }where xi'={h1x(i)..,hr(xi)}
End for
Step 3:Learn a meta class
Learn H based on Dh
Return H

Module 4

SAMPLING

Under sampling: -it means taking the smaller number of majority class (In our case taking a smaller number of Normal transactions so that our new data will be balanced

Oversampling: it means using replicating the data of minority class (fraud class) so that we can have a balanced data

SMOTE: it is also a type of oversampling but in this we will make the synthetic example of Minority data and will give as a balanced data

CONCLUSION

The enlarging rate of intrusions in the network and host machines have badly affected the security and privacy of users. The security feature of intrusion detection using machine learning approach have been appraise in our project, "an ensemble approach has been produced better result over the other approaches" .Examine, divulge that if a method is performing well for detecting an attack, it may not accomplish same for detecting other attacks. In this project, By comparing all the approaches, sampling techniques gives much better result 81% . Hence the relevance of a technique for each type of attack.

REFERENCES

[1] T. F. Lunt, "Ides: An intelligent system for detecting intruders," in Proceedings of the Symposium: Computer Security, Threat and Countermeasures, 1990, pp. 30–45.
[2] A.-S. K. Parham, The State of the Art in Intrusion Prevention and Detection. CRC Press, 2014.
[3] G. Kim, S. Lee, and S. Kim, "A novel hybrid intrusion detectionmethod integrating anomaly detection with misuse detection," Expert Systems with Applications, vol. 41, no. 4, pp. 1690–1700, 2014.
[4] M. Rostamipour and B. Sadeghiyan, "Network attack origin forensics with fuzzy logic," in Computer and Knowledge Engineering (ICCKE),2015 5th International Conference on. IEEE, 2015, pp. 67–72.
[5] N. Hoque, M. H. Bhuyan, R. C. Baishya, D. Bhattacharyya, and J. K.Kalita, "Network attacks: Taxonomy, tools and systems," Journal of Network and Computer Applications, vol. 40, pp. 307–324, 2014.
[6] A. A. Ghorbani , W. Lu, and M. Tavallaee, "Network attacks," in Network Intrusion Detection and Prevention. Springer, 2010, pp.1–25.
[7] S. Selvakani and R. Rajesh, "Genetic algorithm for framing rules for intrusion detection," IJCSNS International Journal of Computer Science and Network Security, vol. 7, no. 11, pp. 285–290, 2007
[8] A. Ahmad and L. Dey, "A k-mean clustering algorithm for mixed numeric and categorical data," Data & Knowledge Engineering, vol. 63, no. 2, pp. 503–527, 2007
[9] A. Chandrasekhar and K. Raghuveer, "Intrusion detection technique by using k-means, fuzzy neural network and svm classifiers," in Computer Communication and Informatics (ICCCI), 2013 International conference on,IEEE,2013,pp.1-7.