International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-4, February 2020

Intrusion Detection System using KDD Cup 99 Dataset

Ch. Aishwarya, N. Venkateswaran, T. Supriya, M. Sreekar, V. Sreeja

Abstract: ***Intrusion Detection System is a vital feature of protecting network infrastructure from unauthorized users or hackers. Intrusion detection system is used to identify several types of malicious activities that could effect the safety of network and to reduce network traffic. Because of faster growth of Internet, networks are growing rapidly in every area of society. As a result, large amount of data is travelling across many networks which may lead to vulnerability of integrity and confidentiality of data. Many Machine learning models are opened up providing new opportunity to classify traffic in network. In quest to select a good learning model, this paper illustrates performance between J48, Naive Bayes and Random forest classification models. The KDD Cup 99 dataset is used for experimental analysis to identify which classification model improves correctness of data and attains highest accuracy.***

Indexed terms: Intrusion Detection, Machine Learning, KDD dataset, Classification models, Naive-bayes, J48, Random Forest, WEKA.

I. INTRODUCTION

Network security is one of the major challenge facing in area of Computer Science. Security attacks are possible because of loopholes in designing of software and hardware. Intrusion Detection System(IDS) aid us to defend against vulnerable attacks. IDS acts as a shield to protect networks from malicious attacks and hackers[1]. IDS's can be classified into two ways, either according to source of the events that they monitor like host events or network events, or according to the method they use to perform detection[2]. In general, there are two detection methods exists, namely signature based detection and anomaly based detection. In signature based detection, all defined patterns are compared with network pattern and the IDS is trained to recognize them. If any defined pattern matches with network pattern then the system is said to be attacked.

However, signature based detection fails to detect new attacks since it will not have defined patterns for new attacks. In anomaly based detection, the normal behaviour of network traffic is defined and any network traffic which deviates from normal behaviour mean the network is under attack. Anomaly based detection can detect new attacks.

Both techniques of IDS also have certain disadvantages. Neither of the techniques have proven to be much better than the other technique.

Signature based detection reduces false alarm rate but it is unable to detect new attacks. Anomaly based detection can detect new attacks but may generate false alarm rates. Intrusion detection could be better if the input data was not manual and to be generated by a system itself by constantly learning like human.

In this case, Machine learning classification models can be used which improves the system over time by learning from new data. Machine learning can be defined as "A computer program is set to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at task in T, as measured by P, improves with experience E." Machine learning techniques can be categorized into supervised or unsupervised learning.

II. METHODOLOGY

A well-recognized KDD Cup 99 dataset was used to check performance analysis of various supervised classification techniques in testing phase.

The KDD Cup 99 dataset is trained and tested by using Naive Bayes, J48, Random forest classification models. A machine learning open source tool named WEKA (Waikato Environment for Knowledge Analysis) was used for implementation. The above classification model attains highest accuracy as outcome.

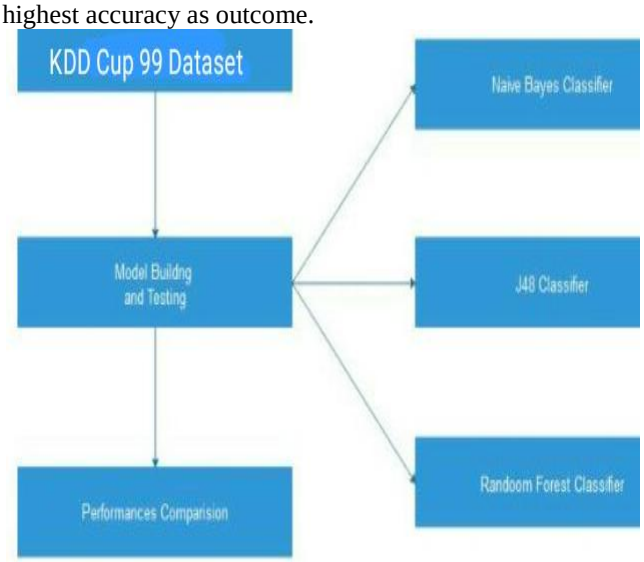


Figure 1 : System Architecture

Revised Manuscript Received on February 28, 2020.
***** Correspondence Author
Ch. Aishwarya*, Department of Computer Science and Engineering, Jyothismathi Institute of Technology and Science, Karimnagar, Telangana, India.
N. Venkateswaran, Associate Professor Department of Computer Science and Engineering, Jyothismathi Institute of Technology and Science, Karimnagar, Telangana, India.
T. Supriya, Department of Computer Science and Engineering, Jyothismathi Institute of Technology and Science, Karimnagar, Telangana, India. **M. Sreekar,** Department of Computer Science and Engineering, Jyothismathi Institute of Technology and Science, Karimnagar, Telangana, India.
V. Sreeja, Department of Computer Science and Engineering, Jyothismathi Institute of Technology and Science, Karimnagar, Telangana, India.

The methodology depicted in Figure 1 describes our machine learning model construction and implementation.

III. RELATED WORK

Some information regarding the classification models used in this paper is provided below:

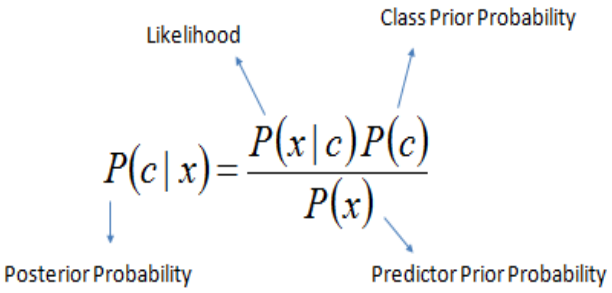
A. J48 CLASSIFIER:

J48 algorithm is an improved version of C4.5 which is a successor of ID3, developed by Ross Quinlan[4]. It supports greedy method and top down approach for Decision tree making. It follows Divide and Conquer strategy. J48 algorithm deals with missing values and noisy data in a dataset. Empty and insignificant branches, overfitting are some of the drawbacks of J48 algorithm. J48 algorithm is mainly used to handle categorical and nominal attributes[5]. Accuracy of J48 classification model is evaluated using entropy. Entropy can be given by,

$$E = -\sum_{j=1}^{k} P_j \log_2 P_j$$

B. NAIVE BAYES CLASSIFIER :

Naive bayes classification model is derived from Bayes theorem. It is mainly used to handle numeric attributes. This algorithm is mainly used for text and document classification[6]. We can easily predict classes using this algorithm and it is very easy to train. It assumes that features are independent of class.Sentiment analysis, Recommendation system and filtering of E-mail spams are some of the application areas of naive bayes algorithm. There are 6 methods of Naive Bayes includes Gaussian naive bayes, Multinomial naive bayes, Complement naive bayes, Bernoulli naive bayes, Categorical naive bayes and out-of-core naive bayes. Probability of Naive-bayes classification model can be given as,



C. RANDOM FOREST :

Random Forest classification model was first stated by Leo Breiman in 2000's. This type of classification models are considered to be one of the most accurate classification model available today.They are very fast and attains high accuracy[7]. Random Forests provides us flexibility to handle large dataset with many number of instances without overfitting.

IV. EXPERIMENTAL ANALYSIS AND RESULTS

As three algorithms namely J48 classifier, Naive bayes classifier, Random forest are being taken for evaluation of Intrusion Detection System. The dataset used for implementation in this paper is KDD cup 99 dataset. This dataset consists of 42 attributes of nominal type consisting of 494020 number of instances.

Our experimental analysis showed that the True positive rate of Random Forest algorithm is 99.99% which means it can detect 99.99% of attacks truly, whereas J48 algorithm showed next highest True positive rate of 99.98% and Naive bayes classifier showed True positive rate of 92.90%.

Correctly Classified Instances	493929	99.9816 %
Incorrectly Classified Instances	91	0.0184 %
Kappa statistic	0.9997	
Mean absolute error	0	
Root mean squared error	0.0038	
Relative absolute error	0.0562 %	
Root relative squared error	2.3707 %	
Total Number of Instances	494020	

Figure2 : Outcome of J48 algorithm

Figure 2 depicts the outcome of J48 classification model with True Positive Rate of 99.98% with 493929 correctly classified instances and 91 incorrectly classified instances.Kappa statistic of the algorithm is 0.9997.

Correctly Classified Instances	458981	92.9074 %
Incorrectly Classified Instances	35039	7.0926 %
Kappa statistic	0.8827	
Mean absolute error	0.0062	
Root mean squared error	0.0766	
Relative absolute error	12.001 %	
Root relative squared error	47.7579 %	
Total Number of Instances	494020	

Figure 3 : Outcome of Naive bayes classification model

Figure 3 depicts the outcome of Naive bayes classification model with True Positive Rate of 92.90% with 458981 correctly classified instances and 35039 incorrectly classified instances attaining Kappa statistic of 0.8827.

Correctly Classified Instances	494018	99.9996 %
Incorrectly Classified Instances	2	0.0004 %
Kappa statistic	1	
Mean absolute error	0	
Root mean squared error	0.0015	
Relative absolute error	0.0381 %	
Root relative squared error	0.9649 %	
Total Number of Instances	494020	

Figure 4: Outcome of Random forest classification model

Figure 4 depicts the outcome of Random forest classification model with True Positive Rate of 99.99%and False Positive Rate of only 0.0004% with 494018 correctly classified instances and 2 incorrectly classified instances.Kappa statistic of the algorithm is 1.

V. CONCLUSION

IDS helps to secure the system in a network connected to a internet from attacks.In this paper, we have chosen J48 classifier, Naive bayes and Random forest algorithms to find which of the three algorithms is more accurate and efficient.The results obtained by the experiments revealed that Random forest performs well for designing and implementation of intrusion detection system. The detection rate of Random forest algorithm nearly tends to 100% of detection rate compared to J48 and Naive bayes classification models.

REFERENCES

1. Ashok Kumar D, Venugopalan Srinivasagopalan Rajan, "The Effect of Normalization on Intrusion Detection Classifiers", International Journal on Future Revolution in Computer Science & Communication Engineering,Volume: 3 Issue: 7, 2017.
2. Uzair Bashir,Manzoor Chacho, "Performance Evaluation of J48 and Bayes Algorithms for Intrusion Detection System", International Journal of Network Security & Its Applications (IJNSA) Vol.9, No.4, July 2017.
3. Vivek Nandan Tiwari, Prof.Satyendra Rathore, Prof.Kailash Patidar, "Enhanced Method for Intrusion Detection over KDD Cup 99 Dataset", International Journal of Current Trends in Engineering & Technology, Volume 2 Issue 2, 2016.
4. Md.Nurul Amin, Md.Ahsan Habib, "Comparison of Different Classification Techniques Using WEKA for Hematological Data", American Journal of Engineering Research, Volume-4, Issue-3,2015.
5. N.Saravanan, V.Gayathri, "Performance and Classification Evaluation of J48 Algorithm and Kendall's Based J48 Algorithm (KNJ48)", International Journal of Computational Intelligence and Informatics, Vol7:No.4, March 2018.
6. Pouria Kaviani, Mrs.Sunita Dhotre, "Short Survey on Naive Bayes Algorithm", International Journal of Advance Engineering and Research Development, Volume 4 Issue 11, November 2017.
7. Gerard Baiu, "Analysis of a Random Forests Model",Journal of Machine Learning Research 13, 2012.

AUTHORS PROFILE

Ch. Aishwarya, is pursuing B.Tech (Computer Science and Engineering) in Jyothishmathi Institute of Technology and Science in Karimnagar, India.

N.Venkateswaran, Working as Associate Professor in CSE department at Jyothishmathi Institute of Technology and Science, Karimnagar. He has 14 years of teaching experience in various engineering colleges and 5 years in industry experience.He has presented papers in many International and National Conferences and his research interests cover Wireless Sensor Networks (WSN), Mobile Ad hoc Network (MANET) and specially Security challenges.

T. Supriya, is pursuing B.Tech (Computer Science and Engineering) in Jyothishmathi Institute of Technology and Science in Karimnagar, India.

M. Sreekar, is pursuing B.Tech (Computer Science and Engineering) in Jyothishmathi Institute of Technology and Science in Karimnagar, India.

V. Sreeja, is pursuing B.Tech (Computer Science and Engineering) in Jyothishmathi Institute of Technology and Science in Karimnagar, India.