INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 5, Issue 2)
Available online at: www.ijariit.com

Classification of attack types for intrusion detection system using machine learning algorithm: Random forest

Dasari Sree Lalitha Chinmayee
dasarilalitha234@gmail.com
Anil Neerukonda Institute of Technology and Sciences, Bheemunipatnam, Andhra Pradesh

C. Visishta
cvisishta.15.cse@anits.edu.in
Anil Neerukonda Institute of Technology and Sciences, Bheemunipatnam, Andhra Pradesh

Garbhapu Navya
gnavya.15.cse@anits.edu.in
Anil Neerukonda Institute of Technology and Sciences, Bheemunipatnam, Andhra Pradesh

Sajja Ratan Kumar
srathankumar.cse@anits.edu.in
Anil Neerukonda Institute of Technology and Sciences, Bheemunipatnam, Andhra Pradesh

ABSTRACT

In the current era of Big Data, a high volume of data is being grown in vast and the speed of generating the new data is accelerating quickly. Machine Learning algorithms are used for such large datasets to teach computers how to reply to and act like humans. In machine learning with the help of generalization ability, the increase in the size of the training set increases the scope of testing. In this paper, we analyze the results of the attacks classified using Intrusion Detection System, and the training time of Random Forest algorithm is measured by increasing the size of the KDD dataset in intervals thereby observing the changes in the final evaluation metrics obtained

Keywords— *Classification, Intrusion detection system, KDD dataset, Evaluation metrics*

1. INTRODUCTION

With the rapid growth in information technology in the past few decades, computer networks are being widely used by industry, business and various fields of human life. Therefore the construction of reliable networks is an important task for network administration. On the other hand, the rapid development of IT produced several challenges to develop reliable networks, which is a difficult task. There are different types of attacks threatening the integrity, availability and confidentiality of computer network. Intrusions are considered as one of the most harmful attacks. An intrusion detection system checks the network traffic and gives alerts if any suspicious activity or attack is found. It also serves as a defence system to protect the information which is stored on various computer platforms. In the case of known attacks administrator can easily judge and process it immediately but it is difficult to judge and process abnormal attacks and the cost of restoration also increases [2]. Although IDS checks for malicious activity by monitoring the data it can raise false alerts [4]. By applying the machine learning techniques we can improve the Intrusion Detection Systems (IDS). Machine Learning Algorithm is widely used in IDS because its capability to classify normal/attack network packets by learning the patterns based on collected data. The authors have conducted an experiment to calculate several performance classifications based on KDD dataset. The Random Forest machine learning algorithm has been implemented, and the time taken to train the dataset is measured. The dataset is divided into a number of divisions in order to observe the level of increase in the evaluation metrics linearly with the increase in the size of the dataset. The Random Forest Algorithm is used for regression and classification problems, it is an ensemble of different decision trees for the same dataset.

2. RELATED WORK

The Big Data revolution has an ability to transform how we live, work, and think by enabling process optimization, empowering insight discovery and also by improving decision making. The realization of this great potential relies on the ability to draw out value form massive data with the help of machine learning and data analytics. As the data is growing rapidly, it is observed that illegal activities such as unauthorized data access, data theft, data modifications and various other intrusion activities are growing rapidly during the last decade. So, deployment and continuous improvement of Intrusion Detection System (IDS) are of greatest importance [6]. The KDD dataset was first publicized by MIT Licoln lans at University of California in 1999. Random Forest Algorithm was aimed at enhancing the tree classifiers based on the concept of forest [1]. The authors figured out the number of trees generated in order to predict the expected outputs. The authors mainly focused on calculating the True Positive and True Negative metrics in order to achieve the highest accuracy rate with Random Forest[8]. The increase in the accuracy rate is observed, together with the other evaluation metrics by

increasing the size of the dataset in intervals. Based on the recent developments and contributions in the networks area, the authors have observed that the training time of the algorithm is not being calculated until date, which has been the motivation for calculating the elapsed time (Training time of the model) for different data samples in the dataset.

3. EVALUATION METRICS

Evaluation metrics are used calculating and observing the performance of the IDS and for comparing the results obtained from the dataset. [3]The performance of the intrusion detection system (IDS) is evaluated by calculating four metric values, Accuracy, Precision, Recall and F-score, out of which accuracy plays a major role and the performance evaluation of the IDS is mainly dependent on accuracy metric.

3.1 Accuracy

This metric is calculated by finding the total number of instances that are correctly predicted as positive cases to the total number of data that is present, the instances are classified into positive or negative cases by calculating the data that are divided into True Positive(TP), True Negative(TN), False Positive(FP), False Negative(FN).True Positives(TP) are the data which are correctly classified as true instances, True Negatives(TN) are the data which are correctly classified as false instances, False Positives (FP) refers to the data that are negative instances but are predicted as positive and False Negative(FN) refers to the data that are positive instances which are predicted as negative. The accuracy rate at the maximum times can be taken as high though there are less number of negative instances which does not play a major role in decreasing the accuracy rate, it is calculated as:

$$\text{Accuracy} = TP+TN/TP+TN+FP+FN$$

3.2 Precision

Precision refers to the total data which are correctly predicted to be positive over the total number of data that are predicted to be positive, by observing the false positive and true positive instances, precision can be calculated as:

$$\text{Precision} = TP / TP+FP$$

3.3 Recall

Recall also known as a True Positive Rate (TPR), sensitivity (SN) or detection rate indicates the total number of instances that are correctly predicted as positive over the total number of actually positive instances present. While detecting the overall positive data in the dataset the recall serves as the main evaluation metric or the best performance indicator of positive data, it is calculated as follows:

$$\text{Recall} = TP / FN+TP$$

Precision and Recall are equally important for calculating the performance of the IDS, each individual is not sufficient for the evaluation of the performance of IDS.

3.4 F-score

F-score is calculated by considering both the metrics of precision and recall equally, the f1 and f2 scores are calculated, in case of f1 both the metrics are treated equally and the value is obtained by substituting 1 in the place of f-beta, in the case of f2 score the recall is considered two times more important than precision.

4. EVALUATION

The procedure of the experiment and the evaluation results obtained are discussed in this section.

4.1 Data Preparation

KDD dataset is well known for benchmarking intrusion detection techniques. The dataset is a massive collection of 9123 KB of data collected over months. [1]The KDD dataset, the authors collected consists of approximately 1,27,426 records, each of consists of 41 features and is labelled either normal or an attack. With exactly one specific attack type and the attacks simulated falls in one of the following four categories, DOS (Denial of service), U2R (used to root attack), R2L (Remote to local attack), and probing attack. Therefore, we set the duration of the data collected a month, so that dataset contains different attack types and has enough data per attack.

4.2 Selection of evaluation metrics and Machine Learning Algorithm

For Intrusion Detection Algorithm it is important to have knowledge on Recall more important than that of precision, so we require F-score. The Random Forest Algorithm was implemented in R version RX64 3.4.1. The PC specifications used in the process of evaluation is Intel core i5 on Windows operating system.

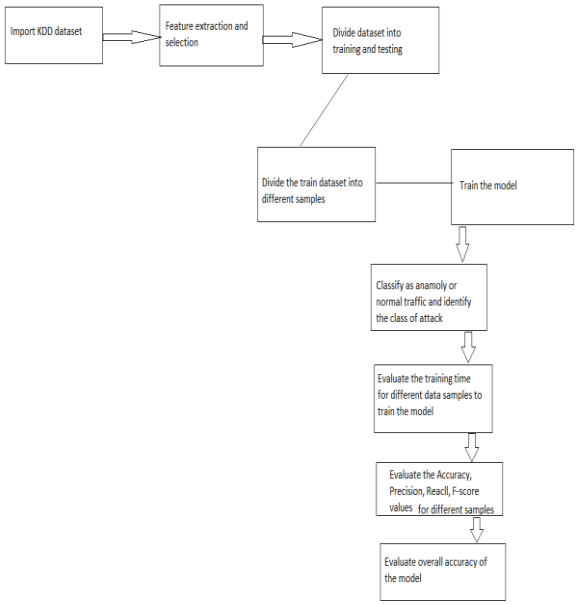


Fig. 1: It describes how the evaluation matrix is calculated from the KDD dataset

4.3 Calculation of Training time

The dataset is divided into subparts and it is subjected to training, with the increase in the size of the dataset, the time taken to train the model will increase sequentially, which is known as training time, this training time is observed.

4.4 Increase in percentage of evaluation metrics

With the increase in the size of the dataset, in particular sequence, the values of evaluation metrics are increased accordingly, and these percentage increase in Accuracy, Precision, Recall, F-score is observed for all data samples is observed.

5. RESULTS

After training the dataset with the algorithm the following results are obtained.

Table 1: Results

No. of data samples in training data	Accuracy	Precision	Recall	F-Score
10240	0.972	0.629	0.822	0.640
20480	0.970	0.639	0.867	0.671
30720	0.972	0.612	0.713	0.639

40960	0.972	0.617	0.766	0.644
51200	0.970	0.617	0.766	0.654
61440	0.968	0.561	0.707	0.573
71680	0.969	0.594	0.726	0.625
81920	0.971	0.637	0.796	0.659
92160	0.969	0.594	0.752	0.610
10490	0.970	0.597	0.777	0.624

Table 2: Training Time

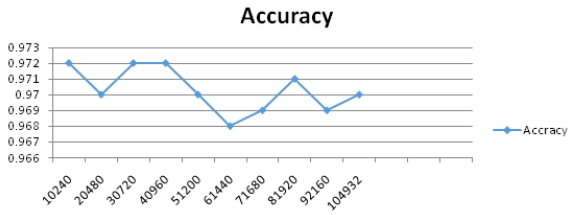
No.of data samples in training data	Elapsed Time
10240	21.07 sec
20480	33.43 sec
30720	45.28 sec
40960	59.54 sec
51200	1.19 min
61440	1.43 min
71680	1.67 min
81920	1.89 min
92160	2.14 min
10490	2.52 min

Table 3: Overall Results

	Accuracy	Precision	Recall	F-score
Normal	0.9737	0.9813	0.9460	0.9634
Dos	0.9319	0.9489	0.9229	0.9355
Probe	0.9644	0.7763	0.8569	0.8146
U2R	0.9801	0.2050	0.5562	0.2996
R2l	0.9977	0.0232	0.1428	0.04
Result	0.9696	0.5869	0.6850	0.6106

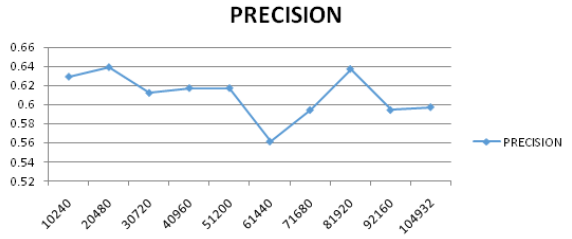
5.1 Graphs

The graphs are plotted for the evaluation metrics calculated as below:



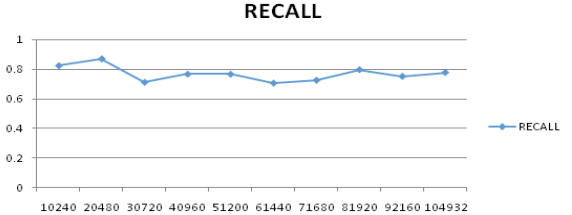
X-axis: No. of data samples in training dataset Y-axis: Accuracy

Fig. 2: Accuracy graph



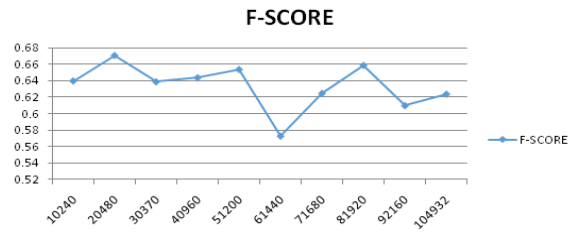
X-axis: No.of data samples in training dataset Y-axis: Precision

Fig. 3: Precision graph



X-axis: No.of data samples in training dataset Y-axis: Recall

Fig. 4: Recall Graph



X-axis: No.of data samples in training dataset Y-axis: F-score

Fig. 5: F-score graph

X-axis: No. of data samples in training dataset Y.axis: training time

Fig. 6: Training time graph

6. CONCLUSION AND FUTURE WORK

The authors have analyzed the class specific detection with the KDD dataset, using the supervised machine learning algorithm Random Forest for IDS and the test data and the training data is constructed for evaluating the performance to detect different types of attacks. The training time of the model is observed with the respective increase in the size of the dataset. The increase in the values of evaluation metrics (Accuracy, Precision, Recall, F-score) by increasing the size of the dataset in steps is observed. From the experiment conducted the authors obtained the results with an accuracy of 96%.

In future will try to investigate the performance of detecting the attack types using graphical interface NVIDIA GPU GEFORCE for the better performance we will implement the same in parallel computation.

7. ACKNOWLEDGEMENT

The authors would like to thank Mr. Sajja Rathan Kumar (assistant professor) dept. of cse for extending his support and guidance in the process of working on this paper.

8. REFERENCES

[1] Evaluation of machine learning algorithm for intrusion detection system "https://arvix.org/ftp/arxiv/papers/1801/1801.02330.pdf"

[2] "http://www.daily.co.kr/news/article.html?no=157416"

[3] Kinam Park, Young or Song, YUN_Gyung Cheong, "Classification of attack types for Intrusion Detection System using a Machine learning algorithm"

[4] Sally, Hassen and Sami Bourouis, "Intrusion Detection alert management for high-speed networks: Current research and applications." Security and Communication Networks 8.18(2015): 4362-4372.

[5] Dataset description and Validation" http://shodhganga.inflibnet.ac.in/bitstream/10603/120011/11/11_chapter3.pdf"

[6] Machine learning for Intrusion detection on public-datasets. https://ieeexplore.ieee.org/document/7726677

[7] The description on random forest algorithm https://www.analyticsvidhya.com/blog/2014/06/introduction-random-forest-simplified/

[8] Review for data classification evaluations https://pdfs.semanticscholar.org/6174/3124c2a4b4e550731ac39508c7d18e520979.pdf

