Impact of Dataset Size and Performance Analysis of IDS using Random Forest Algorithm in 'R' Language

Amit Kumar Mishra, Rakesh Chandra Bhadula, Neha Garg, Deepak Kholiya, V. N. Kala

Abstract: With the advancement of new technologies in today's era, Big Data has shown tremendous growth and popularity. With this exaltation , Big data isn't simply presenting challenge as far as volume yet in addition as far as its high speed generation. New data is fetched extremely fast so it becomes essential to deal with such voluminous data. Machine Learning expedites computers in building models from input data so as to automate decision-making processes. Machine learning algorithms such as "Random Forest" is used with the help of certain datasets to instruct and train computers and also train them to respond like human beings. Selecting an appropriate dataset(size, parameters) plays an important role in providing efficient and effective result. In this paper, an analytical approach is used for IDS i.e. "Intrusion Detection System "where " Random Forest algorithm" is used to analyze the training time by increasing the size of the dataset and observe the impact of frequent changes(size) on various evaluation metrics .Finally performance analysis is carried out and It is observed that the performance of IDS is better and more accurate .

Keywords: Intrusion Detection System, Data set, Evaluation metrics, Machine learning, Random Forest

I. INTRODUCTION

With the fast growth in information technology , computer networks are being widely utilized by industry, business and various domains of human life which in turn produces large amount of information-called as "Big Data" [7] that has to be stored (in electronic devices) and communicated through secured channels in a network. This arouses the construction of reliable networks for various network related domains. On the other hand, the advancement of IT has imposed several challenges to develop reliable networks. There are various categories of attacks undermining 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 defense system to protect the information which is stored on various computer platforms. In this study they identified one of the best algorithm which worked effectively to identify different types of attacks that may occurs in the upcoming future [1][10].

On account of known attacks administrator can without much of a stretch judge and procedure it quickly however it is hard to pass judgment and procedure unusual attacks and the expense of reclamation likewise increments [3]. Although IDS checks for malicious activity by monitoring the data and will raise false alerts [4]. By applying the machine learning techniques we can improve the Intrusion Detection Systems i.e. IDS. Few Machine Learning Algorithms are widely utilized in IDS because it is capable to categorize normal/attack network packets by learning the patterns dependent on the acquired data[8]. The authors have conducted an experiment to calculate several performance classifications based on dataset. The Random Forest machine learning algorithm has been implemented, and the time taken to train the dataset has been measured. The dataset is divided into a number of divisions to observe the increasing level of evaluation metrics linearly with the increasing size of the dataset. The Random Forest Algorithm is utilized for regression and classification problems, it is an ensemble of different decision trees for the same dataset [9].

The Big Data unrest has a capacity to change how we live, work, and think by empowering process optimization, empowering insight discovery and also by improving decision making[7]. The acknowledgment of this extraordinary potential depends on the capacity to draw out value form massive data with the help of machine learning and data analytics. As the data is increasing rapidly, it is observed that illegal activities such as "unauthorized data access"," data theft", "data modifications" and various other intrusion activities are increasing drastically during the last decade. So, deployment and continuous improvement of Intrusion Detection System (IDS) holds great significance [5]. 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 [2][13]. 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[6]. The increase in the accuracy rate is observed together with other evaluation metrics by increasing the size of the dataset at 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.

Revised Manuscript Received on March 20, 2019.

Amit Kumar Mishra, Department of Computer Science & Engineering, Graphic Era Hill, University, Dehradun, India.

Rakesh Chandra Bhadula, Department of Mathematics, Graphic Era Hill, University, Dehradun, India.

Neha Garg, Department of Computer Science & Engineering, Graphic Era (Deemed to be) University, Dehradun, India.

Deepak Kholiya, Department of Agriculture, Graphic Era Hill, University, Dehradun, India.

V. N. Kala, Department of Applied Science, GBPC, Pauri Grahwal, India.

II. EVALUATION METRICS

Evaluation metrics are utilized for calculating and observing the performance of the IDS. The performance of the intrusion detection system (IDS) is evaluated by calculating five metric values; "Accuracy"," Precision", "True Positive Rate (TPR)", "F-score" and "False Positive Rate", out of which accuracy plays a major role and the performance evaluation of the IDS is mainly dependent on accuracy metric[11].

1.1 Selection Of Evaluation Metrics And Machine Learning Algorithm

To design and develop an Intrusion Detection Algorithm, the knowledge about "True Positive Rate" and "Precision" is required and for that we need to calculate "F-score". We have applied Random Forest Algorithm and it was implemented in language R version 3.6.2.

1.2 Change In Percentage Of Evaluation Metrics

With the increment in training sample data set, the value of all related factors like "Accuracy", "Precision", "True Positive Rate", "F-score" are also increasing and "False Positive Rate" and nullity rate are decreasing.

After training the dataset with the algorithm the following results have been obtained:

III. RESULTS

We formulate evaluation metrices for calculating and observing the performance of the intrusion detection system. Such performance of IDS is observed by calculating five metric values: 1) "Accuracy" 2) "Precision" 3) "True Positive Rate (TPR)" 4) "F-score" 5) "False Positive Rate".

Table: 1 Evaluation Of IDS Metric:

No. of Data Samples in Training Data	True Positive (TP)	True Negative (TN)	False Positive (FP)	False Negative (FN)	Accuracy	Precision	True Positive Rate	F-Score	Training Time (Second)	False Positive Rate	Nullity	Nullity Rate
1818	1200	476	110	30	0.923	0.916	0.976	0.945	19.17	0.188	2	0.001
2981	1900	883	115	80	0.935	0.943	0.96	0.951	42.87	0.115	3	0.001
3827	2700	900	130	90	0.942	0.954	0.968	0.961	54.55	0.126	7	0.002
5221	4000	954	160	100	0.95	0.962	0.976	0.969	84.32	0.144	7	0.001
7133	5333	1500	200	90	0.959	0.964	0.983	0.974	124.5	0.118	10	0.001
9249	7000	1900	250	90	0.963	0.966	0.987	0.976	170.82	0.116	9	0.001
11995	9000	2535	350	100	0.962	0.963	0.989	0.976	215.52	0.121	10	0.001
18721	14000	4211	299	200	0.973	0.979	0.986	0.982	257.2	0.066	11	0.001
23982	18000	5455	300	220	0.978	0.984	0.988	0.986	337.67	0.052	7	0
32323	25000	6660	350	300	0.98	0.986	0.988	0.987	425.8	0.05	13	0

An observation is made by calculating these metric values that accuracy plays an significant role for the evaluation of the performance of an intrusion detection system. We calculate the accuracy metric with respect to "True Positive(TP)", "True Negative(TN)", "False Positive(FP)", "False Negative(FN)". The data which is based on true instances is called True Positive(TP), while the data which is related to the false instances is known as True Negatives(TN), False Positives (FP) is the data which refers to the negative instances and a data which is negative instances but predicted as positive is called False Negative(FN). The accuracy is obtained with the help of below expression:

Accuracy =TP+TN/TP+TN+FP+FN

The following graph is plotted from theTable-1, in this graph we demonstrated with the help of sample data that how accuracy is increasing as true positive samples are increasing.

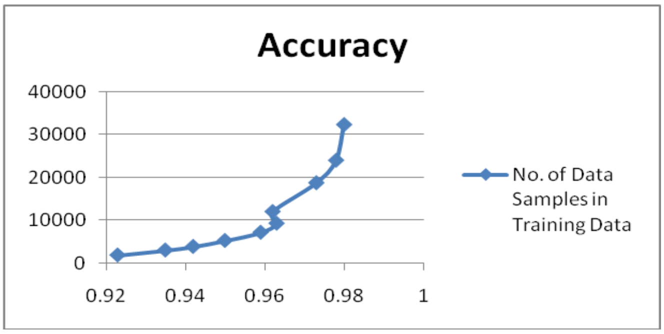

Figure:1 Variation Of Sample Data And Accuracy

Precision is calculated from table-1 with the help of the formula TP / TP+FP, which refers that when true positive samples are increasing, precision is also increasing which can be demonstrated with the help of following graph.

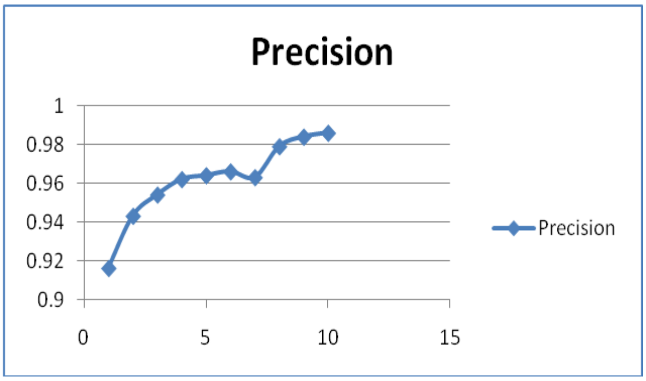

Figure:2 Variation Of Sample Data And Precision

Another very important parameter that is True Positive Rate (TPR) is calculated with the help of the formula TP / FN+TP. True positive metric is increasing when true positive samples increases and false negative decreases. With the help of detection rate ,we may predict that the total number of instances are as positive as the total number of positive instances present, Which can be shown with the help of figure number 3:

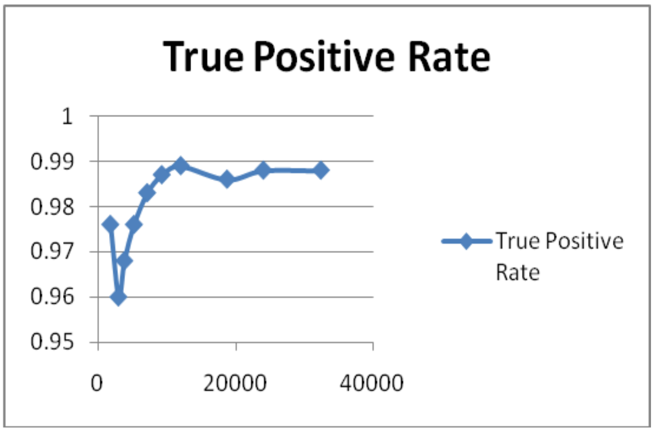

Figure:3 Variation Of Sample Data And True Positive Rate

Both the Precision and Recall are important for evaluating the performance of the Intrusion Detection System.

False positive rate is discussed in figure number:4 which indicates that the performance of intrusion detection system will be better and more accurate if false positive rate will have less value. Technically false positive rate is referred to as type-I error.

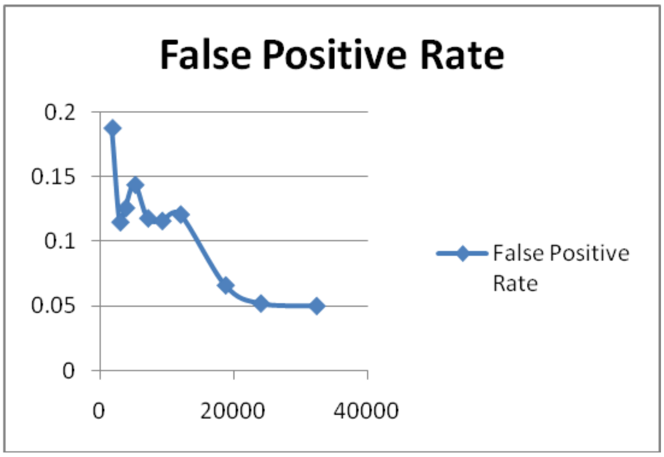

Figure:4 Variation Of Sample Data And False Positive Rate

By taking two metrics precision and True Positive Rate, the two scores F1 and F2 are calculated. F-Score is calculated as the weighted harmonic mean of the precision and true positive rate. Both true positive rate and precision are positive predictor, which shows that when true positive rate and precision will increase simultaneously, F-score will also increase. It means the IDS performance will be more accurate. That is demonstrated in the below figure.

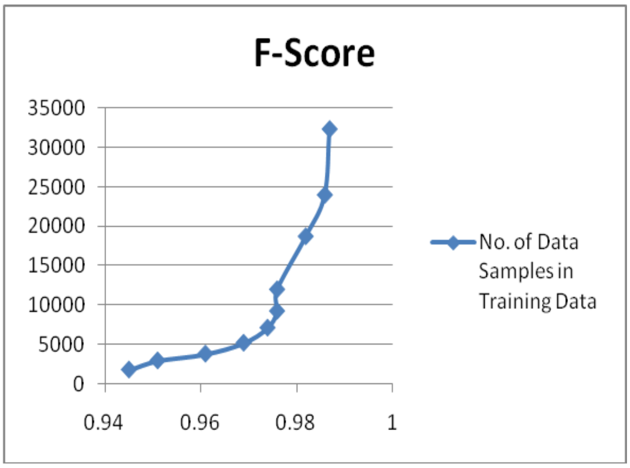

Figure: 5 Variation Of Sample Data And F-Score

The formula that is NULLITY/TP+TN+FP+FN+NULLITY is used to calculate the nullity rate. If the nullity rate increases than the performance of IDS automatically decreases. From table:1 Nullity rate is very less that is why the performance of IDS goes higher and IDS will produce more accurate results.

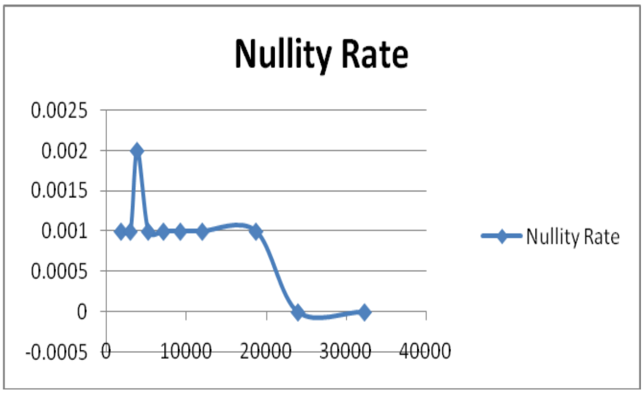

Figure: 6 Variations of Sample Data And Nullity Rate

For evaluation of IDS, training time is one of the most important parameter. With subject to training we divide the dataset into subparts and increased training time with respect to size of the dataset, the total time will increase simultaneously to train the model, this time is observed as training time[12].

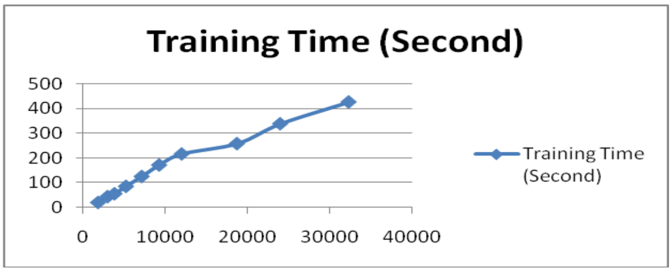

Figure: 7 Variation Of Sample Data And Training Time

IV. CONCLUSION

We use the concepts of machine learning and apply random forest method to analyze the IDS. For analyzing the performance of IDS, we calculate: “Accuracy”, “Precision”, “True Positive Rate”, “F-Score”, “False Positive Rate” and “Nullity Rate”. It is concluded that when the value of data set ascends, the value of evaluation metric’s factors i.e. accuracy, precision, true positive rate, F-Score also ascends and factors like false positive rate and Nullity plummets. It is observed that the performance of IDS is better and more accurate . To demonstrate and discuss these evaluation metrics, we used the Random Forest Algorithm and it was implemented with language ‘R’ and various graphs were also plotted to demonstrate the concepts precisely. From the conducted experiment, we obtained the results with an accuracy rate of 95.6%.

REFERENCES

1. R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat and S. Venkatraman, "Deep Learning Approach for Intelligent Intrusion Detection System," in IEEE Access, vol. 7, pp. 41525-41550, 2019.
2. Evaluation of machine learning algorithm for intrusion detection system, “https://arvix.org/ftp/arxiv/papers/1801/1801.02330.pdf”
3. “http://www.daily.co.kr/news/article.html?no=157416”
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. Machine learning for Intrusion detection on public-datasets. https://ieeexplore.ieee.org/document/7726677
6. Review for data classification evaluations https://pdfs.semanticscholar.org/6174/3124c2a4b4e550731ac39508c7d18e520979.pdf
7. Neelam Singh, Neha Garg, Varsha Mittal,” Big Data – insights, motivation and challenges, in International Journal of Scientific & Engineering Research, Volume 4, Issue 12, December-2013.
8. Yasir Hamid, M. Sugumaran and V. R. Balasaraswathi,”IDS Using Machine Learning - Current State of Art and Future Directions”, British Journal of Applied Science & Technology 15(3): 1-22, 2016
9. Ahmad, M. Basheri, M. J. Iqbal and A. Raheem, "Performance comparison of support vector machine random forest and extreme learning machine for intrusion detection", *IEEE Access*, vol. 6, pp. 33789-33795, 2018.
10. Wanda, Putra. ”A Survey of Intrusion Detection System.” International Journal of Informatics and Computation 1, no. 1 (2020): 1-10.
11. Kumar G ,” Evaluation Metrics for Intrusion Detection Systems - A Study”,International Journal of Computer Science and Mobile Applications (IJCSMA) 2 (11): 11-17, 2014.
12. Cagatay Catal, Banu Diri, “Investigating the effect of dataset size, metrics sets, and feature selection techniques on software fault prediction problem”, Information Sciences Volume 179, Issue 8, 29 March 2009, Pages 1040-1058.
13. Cutler A., Cutler D.R., Stevens J.R. (2012) Random Forests. In: Zhang C., Ma Y. (eds) Ensemble Machine Learning. Springer, Boston, MA