Intrusion Detection System For Big Data Driven Application Using Apache Spark

[1]M. Kranthi Kumar, [2]K. Raja Kumar,

[1]M.Tech Student, [2]Assistant Professor ,

[12]Department of Computer Science And System Engineering,

[12]Andhra University College of Engineering (A), Visakhapatnam , India

mkranthikumar.145@gmail.com

Abstract

The increasing usage of internet based service, the size of network traffic is becoming so large and complex. The conventional processing tool may face difficulties in order to handle this large and complex data, so a fast and efficient intrusion detection is very challenging issue to deal with this data. An intrusion detection system is needed to process this large size of network traffic data to detect the malicious traffic as early as possible. In this paper a big data processing tool Apache Spark is used for processing the large size network traffic data and feature selection is applied to obtain the reduced size dimensionality of data. The malicious traffic data is categorized into DOS(Denial Of Service), R2L(remote to local), U2R(user to root) and probe attacks. Four well-known classification algorithms like Logistics Regression, Random Forest, Naive Bayes and Gradient Boosted Trees are used to find malicious traffic data. In this paper which of these algorithms works well in the context of this big data are identified. The performance of classification algorithms are evaluated in terms of classification time, prediction time, accuracy, recall and specificity.

Keywords: Intrusion Detection System, Feature Selection, Apache Spark, Classification.

1. Introduction

The use of the Internet and computers are increasing, so the security becomes a challenging issue. Different intrusion detection systems are developed to detect the traffic data is either an attack or normal activity[1]. Intrusion detection system is used to monitor system and network activities for detecting malicious activities and produce alert messages to control station[2].

The use of Internet makes life easy but it increases risk of abnormal activity or abuse. Intrusion is to attack a network/host against vulnerable services. These intrusions will gain to the access of personal files, destroy sensitive files, steal or gain unauthorized information etc. Malware like viruses, worms, Trojan horses, root kit, botnet, spyware etc. Security systems like anti viruses, firewalls, Intrusion Detection Systems (IDS) are used to protect computer systems from hackers and crackers.

The important task of any organization is to monitor the network flow and detect any network intrusion which is violating the policies of the organization. So an effective intrusion detection system needed to be developed which is efficient enough to monitor network traffic. there are different other methods like information encryption, access control and intrusion prevention but these are not able to detect all attacks and new attacks[3]. Fast detection of attacks is an important aspect of intrusion detection systems.

In this paper the attacks are categories into 4 major types of attacks like DOS, R2L, U2R, Probe.

Intrusion system are categorized into two different types they are NIDS and HIDS. NIDS stands for Network based Intrusion Detection System and HIDS stands for Host based Intrusion Detection System. There is a basic difference between NIDS And HIDS is that the placing of Intrusion detection system in the organizational environment. In NIDS the IDS placed at nodes within the network so they perform analysis on traffic and alert the administrator if there is an abnormal activities takes place. In HIDS the IDS placed at each and every host or network devices. The source of input data, IDSs can be classified as either network or host based IDS . Network-based systems collect data from network traffic (e.g., packets from network interfaces in promiscuous mode) while host-based systems collect events at the operating system level, such as system calls, or at the application level. Host-based IDS collect high data from the affected system and are not influenced by encrypted network traffic.

For IDS research NSL-KDD, KDD Cup99 datasets are most widely used. However, the dataset was collected in 1999. NSL-KDD Cup 99 dataset is an improved version of KDD dataset and publically available for research. In addition, this dataset is collected from a virtual network environment, which makes it different from the patterns observed in real network systems.

In this paper three well-known feature selection algorithms are used. They are Information Gain, Gain Ratio and Correlation are applied to reduce data set dimensionality. First feature selection applied to identify important features and using those important features and then perform classification to obtain better result.

Data mining schemes plays an important role in intrusion detection system[3-4]. It helps us to obtain patterns from this large and complex network traffic data that helps us to find out the normal or any special kind of attacks. The usage of network and smart devices increasing, so the size of network traffic data is also increasing at large scale. So there is a need to handle this data and produce better results in a small-time.

In order to deal this large and complex network traffic data, a big data processing tool Apache Spark[5-6] is used. In this paper four machine learning classification algorithms[7] such as Logistics Regression, Random Forest, Naive Bayes and Gradient Boosted Trees used for classification of network traffic data. In this paper a multi node setup is done for big data processing.

2. Literature Survey

In this section we review the recent enhancements in the network security for network intrusion detection system.

S Choudhury, A Bhowal [8] discussed improved machine learning algorithms necessary for proper detection of network intrusion. They also compared the performance of various classifiers in WEKA and concluded that Random Forest and Bayes Net are suitable.

Rahul C, S. S. Sambare [9] done a survey on various frameworks or methods which were proposed by researchers, and their use and effectiveness in the field of Intrusion Detection. Along with this the different data mining methods and algorithms are also used in Intrusion Detection.

R. Venkatesan, R.Ganesan, A.Arul Lawrence Selvakumar [10] discussed Data mining methods are capable of extracting patterns automatically and adaptively from a large amount of data. Various methods related to intrusion detection system are studied briefly. This survey paper states the methods and techniques of data mining to aid the process of Intrusion Detection and the frameworks which were developed using these concepts. The concept of intercepting these two different fields, gives more scope for the research community to work in this area.

M Alkasessbeh, Ahmad B. A. Hassant, Ghazi A1-Naymat [14] they collected a new dataset that includes modern types of attack, which were not been used in previous research. The dataset contains 27 features and five classes. A network simulator (NS2) was used in this work, because NS2 can be used with high confidence due to its capability of producing valid results that reflect a real environment. The collected data has been recorded for different types of attack that target the Application and network layers. Three machine learning algorithms (MLP, Random Forest, and Naïve Bayes) were applied on the collected dataset to classify the DDoS types of attack namely: Smurf, UDP-Flood, HTTP-Flood and SIDDOS. The MLP classifier achieved the highest accuracy rate .

M Almseidin, M Alzubi, S Kovacs [18], several experiments were performed and tested to evaluate the efficiency and the performance of the following machine learning classifiers: J48, Random Forest, Random Tree, Decision Table, MLP, Naive Bayes, and Bayes Network. All the tests were based on the KDD intrusion detection dataset.

3. Methodology

In this work a systematic process was followed to produce an intrusion detection system for detection of intruders in the network. A sequence of steps like collecting datasets and make this datasets according to the classification format, than apply feature selection techniques to reduce dataset dimensionality and computational cost of the system. Finally apply classification algorithms to find which algorithm was more accurate in terms of attacks classification and prediction. spark machine learning libraries are used to find normal and abnormal behavior in network activities. The systems performance was measured by evaluation metrics and classification and prediction times of the system. There are 4 modules in the system are explained below.

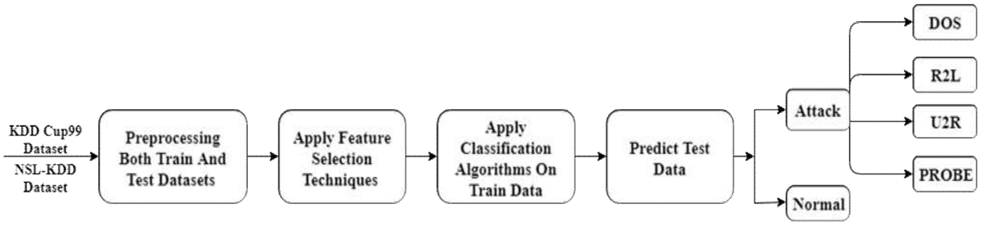

KDD Cup99 Dataset
NSL-KDD Dataset
Preprocessing Both Train And Test Datasets
Apply Feature Selection Techniques
Apply Classification Algorithms On Train Data
Predict Test Data
Attack
Normal
DOS
R2L
U2R
PROBE

Figure 1 : Methodology

3.1 Dataset

In this work both KDD Cup99 data set[17] and NSL KDD are used for network intrusion detection and this are freely available for public and researchers also and it is developed by MIT Lincoln Labs at It is prepared during DARPA 98 intrusion detection evaluation program[15]. This data set consist of 41 attributes which gives appropriate information about network access information. KDDCup99 has both test and train data sets separately. The KDD train data set consist of 4898431 instances and test data set consist of 311029 instances. The NSL KDD train data set consist of 126620 instances and test data set consist of 22850 instances. The data sets consists of four major categories. There are basically 4 categories of attacks taken place in network traffic. Each of these four categories have different types of attacks[18].

- **Pre processing**

  Pre-processing need to be applied in each and every analytical application if it is needed because the quality data that we are giving to system that much quality result It will produce.so Pre-processing is a major step in data mining applications. In present work data set Pre-processing is archived by applying the following.

- **Dataset cleaning**

  In data set cleaning we will take care about missing values and correct the inconsistent data. In our present work missing values not present and removed all null value records in some case null values are filed with mean, median or mode. In order to process inconsistent data we need domain knowledge but whereas in our KDD data set no inconsistent data was present.

- **Dataset transformation**

  In data set transformation nominal attributes are converted to numerical attributes to make classification and prediction easier [19]. In present data set the attributes like protocol type, flag, services, label are normalized. Example all attacks that come under dos are labelled as 1, r2l as 2, u2r as 3,probe as 4 and finally normal as 0, by this we can categorize attack type The table 1 provide clear understanding about data set transformation

- **Dataset Reduction**

  Data set reduction is one of the important step that followed to produce better system. In this process we will remove irrelevant attributes from the data set so that dimensionality was reduced.

3.2 Feature selection

Feature selection is one of the most important step to follow while designing a system, because it is having many advantages like reducing computational cost , increasing system performance and etc. Different feature selection techniques are used to identify the contribution of 41 features in KDD Cup99 Dataset and NSL-KDD Dataset for intrusion detection. Feature selection is applied to reduce data set size without effecting the systems performance. Information Gain, Gain Ratio, Correlation are feature selection algorithms used for feature selection[20, 8]. Importance of feature selection techniques are

- It enables the machine learning algorithm to train faster.
- It reduces the complexity of a model and makes it easier to interpret.
- It try to improves the accuracy of a model.
- It reduces Overfitting.

The KDD Cup99 dataset and NSL-KDD datasets has 41 attributes. From those 41 attribute some attribute are reduced by applying feature selection. Information gain attribute evaluation, gain ratio attribute evaluation and correlation attribute evaluation algorithms are applied and reduced features. the attribute 9, 20 and 21 have no importance and attribute 15, 17, 19, 32, 40 have minimum important in detection of attack. By Observation of datasets the features 7,8,11 and 14 have almost all zero values in dataset. Removing all above features from training and testing set of the dataset we will end up with 29 features, this reduces the size of the dataset. Now the reduced dataset is passed for classification and prediction.

3.3 Classification algorithms

In present work four well-known classification algorithms[7] are used they are 1. Logistic Regression, 2.Random Forest, 3.Naive Bayes, 4. Gradient Boosted Tree. Apache Spark Machine learning libraries used to analyze the performance of system to see the best suitable algorithm for intrusion detection system by comparing with variety of evaluation metrics.

3.4 Evaluation metrics

There are different methods and techniques to evaluate models performance [21]. The best method is Confusion Matrix by this we can easily calculate recall, precision, specificity, and many more. Confusion matrix is a table that is used to measure systems or model performance and It provides visualization of model. But as per our present work above three metrics, classification and prediction times are user to measure the intrusion detection system schemes. The Confusion matrix produces the basic details like True Positive (TP), False Negative (FN), True Negative(TN), False Positive (FP) by using basic knowledge the recall, precision, specificity are calculated and Prediction Time and Classification Time are also calculated.

- Prediction Time: It provides information about how much time is taken to predict entire data set.
- Classification Time: It provides information about how much time is taken to train models.

4. System Overview

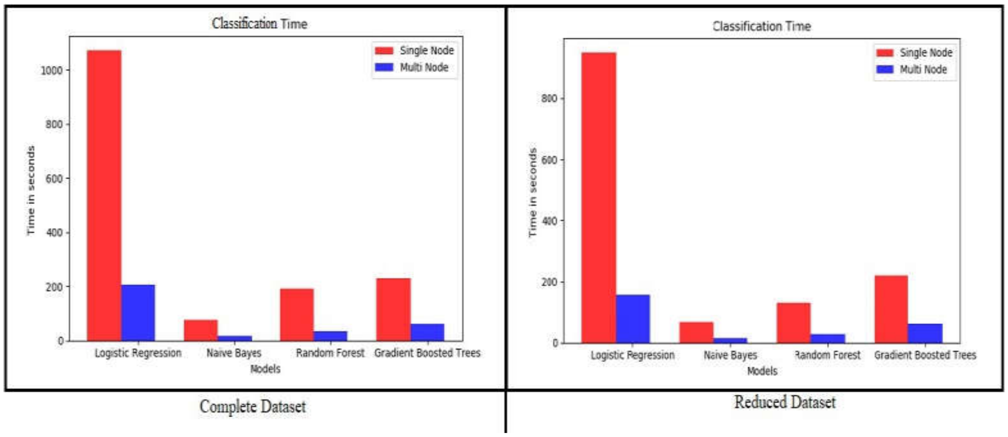
In present work a spark based big data driven application was developed for intrusion detection system for this Ubuntu 18, Anaconda, Apache Spark and Scala are installed in the standalone mode. A multi node setup was created with 3 homogeneous systems of same type are used to check the system performance. The minimum cpu requirement was i5 processor, 500 GB hard disk and 8 GB ram was used. Apache Spark is an open source framework used for big data processing. It supports many languages like java, python, scala to work with it and provides user-friendly API's for handling jobs and writing queries. In Present work Apache Spark was used because it is 100 times faster than other big data processing frameworks like Hadoop and Storm.

5. Results And Discussions

In this experiment, the feature selection algorithms are information gain, gain ratio, and correlation are used to produce the feature set. By applying this feature selection techniques the feature set was reduced to 29 attributes. Now performance analysis of classification algorithms are analyzed. We compare the two data set that are widely used for intrusion detection system. The two datasets are KDD Cup99 and another one is NSL-KDD dataset this are publicly available for the research. So in this work will compare the two dataset available. In present work attacks are categorized to four major type of attacks they are DOS, R2L, U2R, Probe. The main moto is to find out best model for intrusion detection system by different evaluation metrics that are explained in the methodology. The existing system take more time when compared with present system.

5.1 Analysis of KDD Cup 99 Dataset

The below figure 2 and figure 3 are the classification and prediction time of KDD Cup99 dataset. The table 1 and table 2 provides complete information about each and every classifiers efficiency along with attacks classification accuracy. Table 1 provides complete information about dataset without feature selection, whereas table 2 provide complete information about dataset with feature selection. It is observed that naïve bayes classifier works well in terms of classification time and prediction time. In terms of accuracy it will fail. Random forest is the next best in terms of classification time and prediction time and obtain a accuracy 92%. The classification time of complete data set without feature selection in logistic regression with around 1000 seconds is very high and naïve bayes classifier with around 60 seconds is recorded. In multi node also same thing is observed. Whereas in reduced dataset also same thing is observed, but in the context of accuracy the naïve bayes failed, but random forest and logistic regression classifiers achieve the same accuracy but in terms of classification and prediction time the random forest classifiers is the best one.



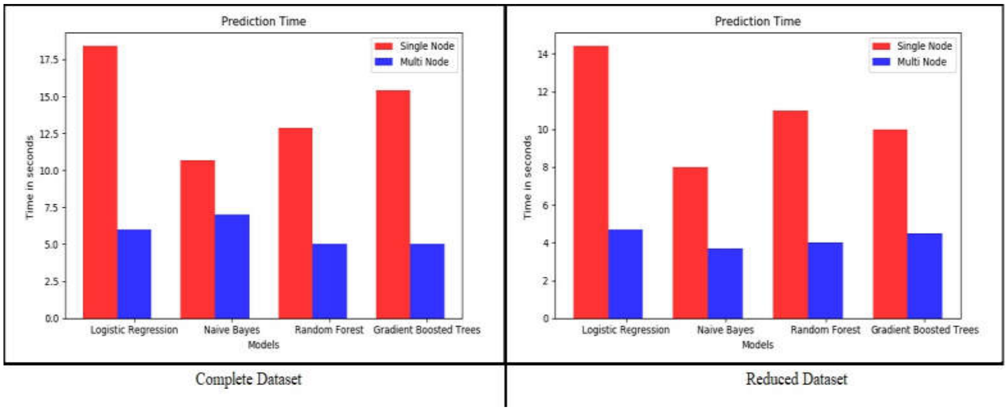
Complete Dataset

Reduced Dataset

figure 2: classification times of both complete and reduced KDD dataset

5.2 Analysis of NSL-KDD Dataset

The below figure 4 and figure 5 are the classification and prediction time of NSL KDD dataset. The table 1 and table 2 provides complete information about each and every classifiers efficiency along with attacks classification accuracy. Table 3 provides complete information about dataset without feature selection, whereas table 4 provide complete information about dataset with feature selection. It is observed that naïve bayes classifier works well in terms of classification time and prediction time. In terms of accuracy it will fail. Random forest is the next best in terms of classification time and prediction time and obtain a accuracy 71%. The classification time of complete data set without feature selection in logistic regression with around 17.5 seconds is very high and naïve bayes classifier with around 14 seconds is recorded. In multi node also same thing is observed. Whereas in reduced dataset also same thing is observed, but in the context of accuracy the naïve bayes failed, but random forest and logistic regression classifiers achieve the same accuracy but in terms of classification and prediction time the random forest classifiers is the best one. It is observe the in the prediction time also they are same.



Complete Dataset

Reduced Dataset

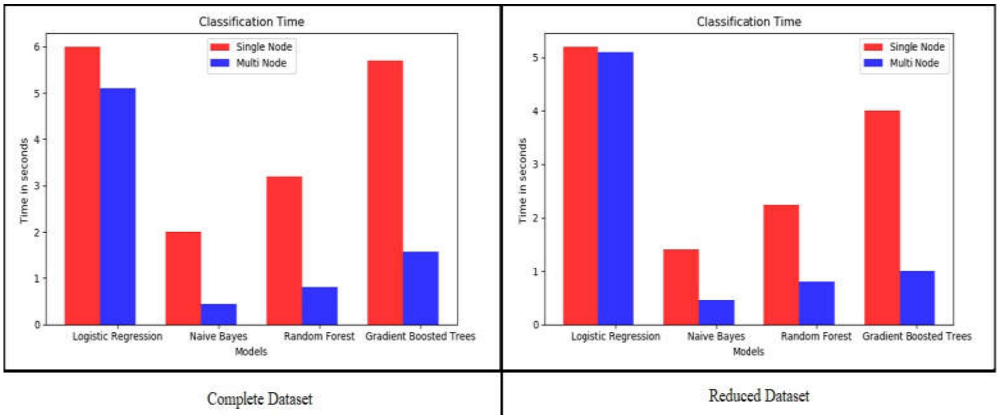
figure 3: prediction time of both complete and reduced KDD dataset

Model	Accuracy	Classification Time		Prediction Time		Evaluation Metrics					
		Single Node	Multi Node	Single Node	Multi Node	Metrics	Normal	DOS	R2L	U2R	Probe
Logistic Regression	92	1073	207	18.42	6	Precision	71	99	91	40	83
						Recall	98	97	72	0	0
						Specificity	90	98	99	99	99
Naive Bayes	83	77.2	17	10.682	7	Precision	79	91	0	17	0
						Recall	61	96	0	3	1
						Specificity	96	74	98	99	96
Random Forest	92	193.2	34	12.84	5	Precision	72	99	33	17	0
						Recall	99	98	4	0	0
						Specificity	90	99	99	100	100
GB Tree	91	230.5	60	15.4	5	Precision	70	98	86	47	66
						Recall	99	96	0	0	0
						Specificity	90	94	100	100	100

Table 1: Analysis Of KDD Cup 99 Dataset without feature selection

Model	Accuracy	Classification Time		Prediction Time		Evaluation Metrics					
		Single Node	Multi Node	Single Node	Multi Node	Metrics	Normal	DOS	R2L	U2R	Probe
Logistic Regression	92	950	158	14.42	4.7	Precision	71	99	91	40	83
						Recall	98	97	72	0	0
						Specificity	90	98	99	99	99
Naive Bayes	80	68.76	15	8	3.7	Precision	79	91	0	17	0
						Recall	61	96	0	3	1
						Specificity	96	74	98	99	96
Random Forest	92	129.49	29	11	4	Precision	72	99	33	17	0
						Recall	99	98	4	0	0
						Specificity	90	99	99	100	100
GB Tree	91	220.5	63	10	4.5	Precision	70	98	86	47	66
						Recall	99	96	0	0	0
						Specificity	90	94	100	100	100

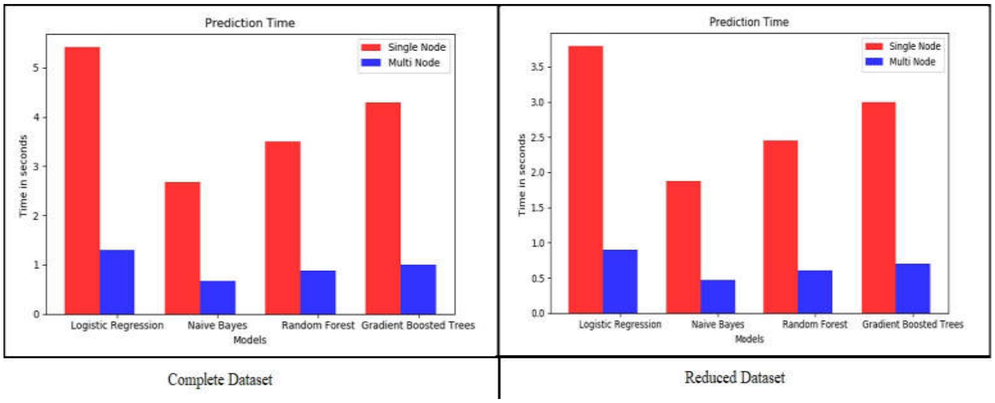
Table 2: Analysis Of KDD Cup 99 Dataset with feature selection



Complete Dataset

Reduced Dataset

Figure 4: Classification time of both complete and reduced NSL KDD dataset



Complete Dataset

Reduced Dataset

Figure 5: Prediction time of both complete and reduced NSL KDD Dataset

Model	Accuracy	Classification Time		Prediction Time		Evaluation Metrics					
		Single Node	Multi Node	Single Node	Multi Node	Metrics	Normal	DOS	R2L	U2R	Probe
Logistic Regression	73	6	5.1	5.42	1.3	Precision	64	90	74	61	66
						Recall	93	78	59	0	2
						Specificity	60	96	97	99	99
Naive Bayes	32	2	0.45	2.682	0.67	Precision	78	42	0	23	0
						Recall	61	96	0	3	1
						Specificity	9	81	0	9	13
Random Forest	68	3.2	0.8	3.5	0.875	Precision	59	95	85	23	0
						Recall	98	63	51	0	0
						Specificity	48	98	98	100	100
GB Tree	70	5.7	1.57	4.3	1	Precision	77	63	0	0	0
						Recall	94	91	0	0	0
						Specificity	79	73	100	100	100

Table 3: Analysis Of KDD Cup 99 Dataset without feature selection

Model	Accuracy	Classification Time		Prediction Time		Evaluation Metrics					
		Single Node	Multi Node	Single Node	Multi Node	Metrics	Normal	DOS	R2L	U2R	Probe
Logistic Regression	71	5.2	3.57	3.794	0.91	Precision	62	90	78	50	62
						Recall	93	74	59	0	1
						Specificity	57	95	98	99	99
Naive Bayes	38	1.4	0.31	1.8774	0.469	Precision	79	44	0	21	0
						Recall	5	81	0	9	13
						Specificity	98	48	99	95	69
Random Forest	71	2.24	0.56	2.45	0.61	Precision	62	93	78	0	0
						Recall	97	73	45	0	0
						Specificity	55	97	98	100	100
GB Tree	70	4	1	3	0.7	Precision	74	66	78	0	0
						Recall	94	90	0	0	0
						Specificity	75	77	100	100	100

Table 4: Analysis Of KDD Cup 99 Dataset with feature selection

3.2 Comparison Of KDD Cup 99 dataset with NSL-KDD Dataset

The KDD Cup 99 dataset is best in terms of all evaluation metrics, but NSL-KDD dataset is also works up to the mark, but the less number of instances is also the factor which influence accuracy. Very less recall values for U2R and Probe in the both KDD and NSL-KDD datasets are record due to unbalanced dataset. So the recall value are very high for normal and dos attacks.in terms of specificity the random forest, logistic regression and gradient boosted trees works well. In NSL-KDD Dataset the accuracy was 71% is recorded, the less number of instances is also one of the factor which influence the accuracy.

6. Conclusion And Feature Scope

The present work is to find algorithms which can accurately classify the records and at the same time take less time for classification as well as predict. Random Forest gave the best performance with respect to all measures accuracy, recall and specificity. It displayed approximately 92% accuracy in just a matter of 130 seconds of training time. Although Naïve Bayes has the least training time amongst all the methods, but compared to all the classifiers a huge difference is noticed in specificity. Logistic Regression, GB Tree give results almost similar to all the measures, that is their values are nearly the same in terms of specificity, sensitivity and accuracy. However, their training time differs by a significant amount. Our future work will be focused on enhancing the accuracy and other measures of these particular algorithms using different technologies. By proposing the new effective feature selection method, classification algorithm and the construction of efficient data set will give us the accurate and best results while processing the data through IDS systems. We will try to implement real time intrusion detection system with new algorithms. The KDD dataset results best in all aspects, whereas NSL KDD dataset also performance best enough with less records.

References

[1] Huang M-Y, Jasper RJ, Wicks TM, "A large scale distributed intrusion detection framework based on attack strategy analysis", Computer Network, pp. 65–75, 1999.

[2] Tan Z, Nagar UT, He X, Nanda P, Liu RP, Wang S, Hu J, "Enhancing big data security with collaborative intrusion detection", IEEE Cloud Computer, pp. 27–33, 2014.

[3] Susan M. Bridges, and Rayford B. Vaughn, "Fuzzy Data Mining and Genetic Algorithms Applied to Intrusion Detection", Proceedings of the National Information Systems Security Conference (NISSC), Baltimore, MD, October, 2000.

[4] Ho C-Y et. al., "Statistical analysis of false positives and false negatives from real traffic with intrusion detection/prevention systems", IEEE Communication Magazine, pp. 46–54, 2012.

[5] Apache Spark™ - Lightning-Fast Cluster Computing, http://spark.apache.org/

[6] Holden Karau et al." Learning Spark", Published by O'Reilly Media, Inc, 2015

[7] C.C. Aggarwal, Data Classification: Algorithms and Applications, CRC Press,2014

[8]. S. Choudhury and A. Bhowal, "Comparative Analysis of Machine Learning Algorithms along with Classifiers for Network Intrusion Detection", International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), 2015, pp.89-95.

[9]. Mr.R. Chinchore and Prof.S.S. Dambare, "A Survey of Techniques for Intrusion Detection System", International Journal of Research in Computer and Communication Technology, Oct-2013,Vol-2,Issue-10, ISSN(Online) 2278-5841, ISSN (Print) 2320-5156.

[10]. R. Venkatesan, R. Ganesan and A.A.L. Selvakumar, "A Comprehensive Study in Data Mining Frameworks for Intrusion Detection", International Journal of Advanced Computer Research , Dec-2012, Vol-2,Number-4,Issue-7, ISSN(print): 2249-7277 ISSN (online): 2277-7970.

[11]. Preeti Aggarwal, Sudhir Kumar Sharma, "Analysis of KDD Dataset Attributes - Class wise For Intrusion Detection". 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015).

[12]. W. Alsharafat, "Applying artificial neural network and extended classifier system for network intrusion detection." International Arab Journal of Information Technology (IAJIT), vol. 10, no. 3, 2013.

[13]. N. Bhargava, G. Sharma, R. Bhargava, and M. Mathuria, "Decision tree analysis on j48 algorithm for data mining," Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, no. 6, 2013

[14]. M. Alkasassbeh, G. Al-Naymat, A. B. Hassanat, and M. Almseidin, "Detecting distributed denial of service attacks using data mining techniques," International Journal of Advanced Computer Science & Applications, vol. 1, no. 7, pp. 436–445.

[15]. A. Lazarevic, L. Ertoz, A. Ozgur, J. Srivastava& V. Kumar, "A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection", Proceedings of Third SIAM Conference on Data Mining, San Francisco, May 2003.

[16]. G. V. Nadiammai and M. Hemalatha, "Perspective analysis of machine learning algorithms for detecting network intrusions," IEEE Third International Conference on Computing Communication & Networking Technologies (ICCCNT), Coimbatore, India, 2012, pp. 1-7.

[17]. KDD Cup 1999 Data: The third international knowledge discovery and data mining tools competition dataset , Avaliableonline:http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

[18]. M Almseidin, M Alzubi, S Kovacs, M Alkasassbeh, "Evaluation of Machine Learning Algorithms for Intrusion Detection System", SISY 2017.

[19]. D Gupta, S Singhal, S Malik, A Singh, " Network Intrusion Detection System Using various data mining techniques", 2016.

[20]. Jiawei Han, Micheline Kamber, Jian Pei, "Data Mining Concepts and Technologies", 3rd Edition .

[21]. M. Kulariya, P Saraf, R Ranjan, P. Gupta, "Performance Analysis Of Network Intrusion Detection Schemes Using Apache Spark", International Conference On Communication and Signal Processing, April 6-8, 2016.