Machine Learning Approach for the Network Traffic Classification using Voting Classifier

Roshani Singh[1], Mr. Saurabh Singh[2]
[1]*Research Scholar,* [2]*Head of Department*
[12]*Buddha Institute of Technology,GIDA,Gorakhpur, India*

Abstract- When two or more computers are connected with each other with the aim of achieving particular benefit, a network is formed. Based on the requirements of users involved in these networks, the information is forwarded or exchanged across the systems. Internet is one of the massive networks where very huge data is being transferred on daily basis. In the networked data, certain interconnected entities generate inferences. For example, for the interconnected web pages hyperlinks are available, the research papers have references and the conceivable terrorists can be linked through phone calls and accounts available in traced conversations. With increasing use of networks, intrusions across the Internet have become a major threat in our world. This research study has been influenced by the different intrusion threats on internet and the ways to detect them. In this research we have studied and analyzed the famous network traffic data -NSL KDD dataset and its various features. The proposed model is a hybrid of Logistic Regression and K-nearest neighbor classifier combined together using voting classifier which aims at classifying the data into malicious and non-malicious with more accuracy than existing methods.

Keywords- Machine Learning, KDD, KNN, LR, Voting

I. INTRODUCTION

The demand of an intrusion detection system in various applications has increased in the recent years since huge amount of data is available to be stored and processed every day. The networking systems are generating huge amount of data by monitoring the surroundings of applications in which they are deployed. Any kinds of suspecting behaviors are detected by the devices. Any kinds of vulnerabilities in any computer network can be found by an intruder that aims to harm the users using that device. For preventing the entry of intrusions, the best solution is to protect the system or its resources [1]. Any activity that attempts to trigger an event due to which the system's security is compromised is called as an intrusion. Either internally or externally, the intrusions might occur in any system [2]. Any kind of illegal activity or fraud information that makes a computer hazardous can be considered as intrusion. An intrusion detection system is the method in which the events being performed in a computer network can be monitored and analyzed so that it becomes easy to recognize the security problems. An intrusion detection systems acts as an alarm and any kinds of violations in the system are identified by it. Even if there are false messages in messages, mails or video sounds, they can be alerted by the systems. A tool that acts as a guard such that the system can be secured against any kinds of intrusions or attacks is called intrusion detection system [3]. To check the attack scenarios and provide required support for defense management are the important objectives of IDS. Today, in networking, almost all the applications are using IDS systems. IDS can be used to detect any malicious activities which cannot be detected by a common firewall. Against the sensitive services, computer applications and other regions, attacks are possible in the computer systems. Data driven attacks are possible in computer applications, network attacks in sensitive services and unauthorized logins in case of sensitive files are faced due to intrusions [4].

Completely accurate detection cannot be ensured by IDPSs only. False positives and false negatives are generated by all these kinds of systems. For reducing the false negatives and increasing the false positives, several organizations choose to tune IDPSs [5]. Thus, the false positives can be differentiated from true malicious events by providing necessary additional analysis of resources. The aim is to reduce the false negatives and increase the false positives. The features which compensate for the usage of common evasion approaches are offered by most of the IDPSs [6]. Thus, the appearance of this method is not affected but the format or timing of malicious activity to alter the appearance is modified such that the detection can be avoided. Either in separated or integrated manner, the multiple detection methodologies are used in most of the IDPs such that more broad and accurate detection can be provided [8]. The primary classes of detection methodologies are as follows:

a. Signature-Based Detection: For the analysis of potentially unwanted traffic, the signature-based detection can be used by an IDS based on the known traffic data. The configuration of this type of detection is fast and easy. In a signature based IDS, an attack can slightly be modified by an attacker such that this system does not recognize it. However, the accuracy of this method can be high even with its limited detection capability. The threat signatures and the observed events are compared for identifying the incidents in these methods. The

known threats can also be detected effectively however, the unknown threats and several variants of known threats are detected by this approach is highly ineffective manner. The state of complex communications cannot be traced and understood in these approaches so most of the attacks that include multiple events cannot be detected here [9].

b. Anomaly-Based Detection: The system that directly looks into the network traffic and identifies the incorrect, invalid or abnormal kind of data is the anomaly-based detection system. The unwanted traffic that is not known specifically can be detected through this method. For example, the detection that an IP packet is malformed can be done through this system. It only detects if the packet is anomalous and does not detect the particular method in which it is malformed [10].

II. LITERATURE REVIEW

AltyebAltaher, (2017) proposed a hybrid mechanism through which the websites as Legitimate, Suspicious, or Phishing websites can be categorized. There are two stages used in this proposed algorithm to generate this hybrid method in which the KNN and SVM classifiers are combined [11]. The KNN is applied in the initial stage which is effective and robust to the noisy data. In the second stage, another powerful classifier is applied which is known as SVM. The effectiveness of SVM algorithm is improved by the proposed algorithm when the simplicity of KNN is integrated. Evaluations are performed by conducting simulation experiments and it is seen through the outcomes that in comparison to other approaches, the accuracy of proposed approach is highest which is 90.04%.

Amol Borkar, et.al (2017) presented a survey of theInternal-IDS and IDS in which the real time based data mining and forensic techniques algorithms were applied [14]. In support of intrusion detection, different data mining techniques were proposed for cyber analytics. Based on the studies of different techniques presented by different authors, this research presented the manners in which the intruder could be detected. The survey presented in this paper helped in drawing the conclusion. The accuracy and detection rate were improved up to 95% by applying the proposed technique in comparison to the existing techniques which provided around 90% of accuracy and detection rate.

Jayshree Jha, et.al (2013) proposed a research that was based on two important contributions. In the initial contribution, the intrusion detection performed using SVM was reviewed in this paper along with the other technologies proposed by different authors [12]. Further, to detect intrusion, the best feature was chosen by proposing a novel method in the second contribution. To select the relevant features, a hybrid approach was proposed in which the filter and wrapper models were combined. The performance and detection accuracy of SVM based detection model were increased by reducing the dataset. Furthermore, with the reduction in theset of feature, it is also possible to reduce the training and testing time.

L.Dhanabal, et.al (2016) performed an analysis of the KSL-KDD dataset. The anomalies in network traffic patterns were detected by studying the effectiveness of different classification algorithms [13]. For generating anomalous network traffic, the relationship of protocols available in commonly used network protocol stack was analyzed with the attacks used by intruders that generated the anomalous network traffic. The data mining tool WEKA was used to perform analysis using the classification algorithms. Several facts that bonded between the protocols and network attacks were exposed in this study

Yi Yi Aung, et.al (2017) analyzed the comparison among hybrid data mining methods and single method in this study [15]. Showing that the usage of hybrid data mining methods can minimize the time complexity of system in comparison to single method is the purpose of proposed method. The KDD'99 data set was used to perform verification of proposed model. The model training time of system was reduced by applying the hybrid methods which included K-means and Projective Adaptive Resonance Theory. Further, the accuracy of detections was maintained in this paper.

Farid Lawan Bello, et.al (2015) analyzed several intrusion detection classifier models on the basis of detection strategies and the implementation techniques. The disadvantages of single classifier models were also reviewed here. ForIDS, the need to design a hybrid intelligent approach was discussed here [15]. This study conducted a comparative analysis of existing hybrid models. To detect the efficiency and the technique used for reducing the time such that the classifiers are not retained again at the time of arrival of new data entry, theexisting hybrid models were compared and the disadvantages were checked for each approachindividually against the proposed hybrid SVMs. Due to the detection accuracy of 94.86%, the performance of CSVAC model was considered to be better in comparison KDD99. Both training and testing phase were implemented to provide an edge over its counterparts in the survey. So, as the training and testing are done simultaneously, it is important to detect real time.

Tahir Mehmood, et.al (2016) compared the various supervised algorithms which could be applied in anomaly-based detection techniques for providing improvements [16]. As a benchmark dataset used for anomaly-based detection techniques, the algorithms were applied on KDD99 dataset. It was seen through the simulation results that for every class of KDD99

dataset, no algorithm provided high detection rate. Further, the true positive, false positive and precision were the three performance metrics used to compares the evaluation results and determine the improvement levels of outcomes.

III. RESEARCH METHODOLOGY

NSL-KDD dataset classification method is performed to categorize the traffic against malicious or non-malicious. This technique helps in predictive the malicious activities of active users. To categorize the network using proposed methodology, three important steps are applied. In the initial step, data preprocessing has been done. As part of pre-processing , scaling of numerical data is performed to achieve zero means and one standard deviation. Encoding of categorial data is done to transform the data into numerical form. Next step is sampling of data and feature extraction based on recursive feature elimination. Then we train the model applying Logistic regression and KNN combined in Voting classifier. Logistic regression regularizes the train data without requiring any tuning and scaling of input features and outputs well-calibrated predicted probabilities. K-NN algorithm scans through complete dataset to find out nearest neighbors based on K-number without making any assumptions about the characteristics of the features in data. Votingalgorithm performs both Logistic regression and K-NN individually and present their probabilities to the voting classifier. This classifier then averages the input and give the class with the highest probability. After training, the last step is to predict the model using test dataset.

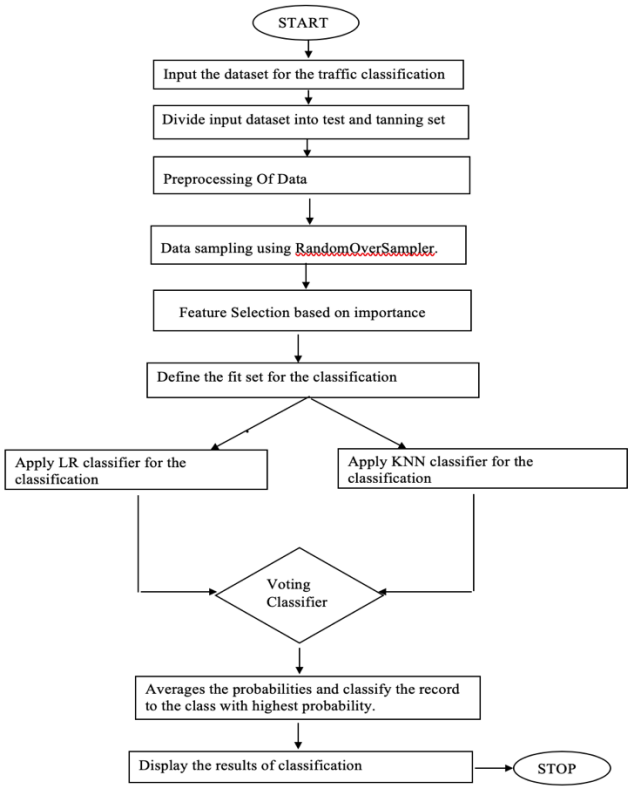

Fig.1: Proposed Flowchart

IV. EXPERIMENTAL RESULTS

The proposed research is implemented in Python and the results are evaluated by comparing proposed and existing techniques.

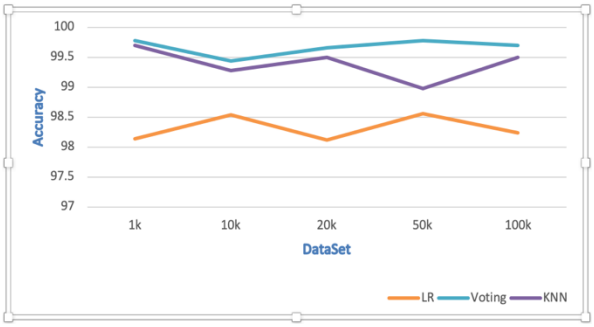

Fig.2: Accuracy Comparison in training phase

The proposed hybrid algorithm performs better in training phase compared to KNN and LR individually. From the prediction phase, we can conclude that the more we train the data ,the better algorithm performs in evaluation phase. Results from evaluation phase on test dataset can be seen in fig:3.

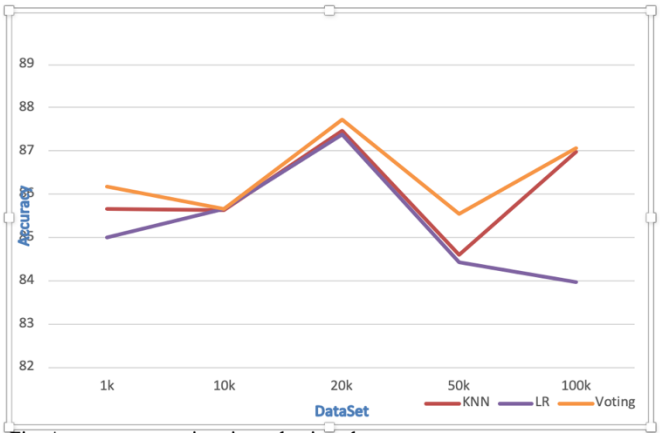

Fig.3: Accuracy Comparison in training phase

Above diagram depicts the comparison of proposed hybrid (LR+KNN) algorithm with KNN and LR individually with different instance run of data sizes. Thus, we can conclude that Voting algorithm gives better precision contrasted with KNN and LR independently. One thing which is worth noticing here is ,Voting performs better with large dataset. As we can point out in above comparison that with 10k data all three algorithms are giving approximately same accuracy relative to each other. But as the dataset increases Voting starts performing better than KNN and LR which we can see above with 20k,50k and 100k datasets.

V. CONCLUSION

The research done is quite useful in understanding how Voting classifier outperforms other classifiers such as LR and KNN[32] , the study also shows the importance of training the model , to improve the overall accuracy of the system. There are various classifiers which can be combined and used in Voting Classifier , however as part of this research , we have used KNN and logistic regression classifier. With this classifier we have observed the more we train the data ,the better algorithm performs in evaluation phase.

VI. REFERENCES

[1]. Amrita, Kiran Kumar Ravulakollu, "A Hybrid Intrusion Detection System: Integrating Hybrid Feature Selection Approach with Heterogeneous Ensemble of Intelligent Classifiers", International Journal of Network Security, Vol.20, No.1, PP.41-55, Jan. 2018

[2]. Bayu Adhi Tama and Kyung-Hyune Rhee, "Performance evaluation of intrusion detection system using classifier ensembles", Int. J. Internet Protocol Technology, Vol. 10, No. 1, 2017

[3]. M. Paz Sesmero, Agapito I. Ledezma and Araceli Sanchis, "Generating ensembles of heterogeneous classifiers using Stacked Generalization", WIREs Data Mining KnowlDiscov 2015, 5:21–34

[4]. Necati DEMIR, Gokhan DALKILIC, "Modified stacking ensemble approach to detect network intrusion", 2018, Turkish Journal of Electrical Engineering & Computer Sciences, 26: 418-433

[5]. Nanak Chand, Preeti Mishra, C. Rama Krishna, Emmanuel Shubhakar Pilli and Mahesh Chandra Govil, "A Comparative Analysis of SVM and its Stacking with other Classification Algorithm for Intrusion Detection", 2016, IEEE

[6]. M. Mazhar, U. Rathore, "Threshold-based generic scheme for encrypted and tunneled Voice Flows Detection over IP Networks", Journal of King Saud University Computer and Information Sciences, vol. 27, pp. 305–314, 2015.

[7]. Muhammad Shafiq, Xiangzhan Yu, Asif Ali Laghari, Lu Yao, N abin Kumar Karn, F oudilAbdessamia, "Network Traffic Classification Techniques and Comparative Analysis Using Machine Learning Algorithms", 2016 2nd IEEE International Conference on Computer and Communications, vol. 8, pp. 2451-2455, 2016.

[8]. JaiswalRupeshChandrakant, LokhandeShashikant. D., "Machine Learning Based Internet Traffic Recognition with Statistical Approach", 2013 Annual IEEE India Conference (INDICON), vol. 7, pp. 121-126, 2013.

[9]. RiyadAlshammari, A. NurZincir-Heywood, "Identification of KDD encrypted traffic using a machine learning approach", Journal of King Saud University – Computer and Information Sciences, vol. 27, pp. 77–92, 2015.

[10]. MazharRathore, Anand Paul, Awais Ahmad, Muhammad Imran, Mohsen Guizani, "High-Speed Network Traffic Analysis: Detecting KDD Calls in Secure Big Data Streaming", 2016 IEEE 41st Conference on Local Computer Networks, vol. 7, pp. 595-598, 2016.

[11]. AltyebAltaher, "Phishing Websites Classification using Hybrid SVM and KNN Approach", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 6, 2017

[12]. Jayshree Jha, Leena Ragha, "Intrusion Detection System using Support Vector Machine", 2013, International Journal of Applied Information Systems (IJAIS), Foundation of Computer Science FCS, New York, USA International Conference & workshop on Advanced Computing

[13]. L.Dhanabal, Dr. S.P. Shantharajah, "A Study on **NSL**-KDD Dataset for Intrusion Detection System Based on Classification Algorithms", 2016, International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 6

[14]. Amol Borkar ; AkshayDonode ; Anjali Kumari, "A survey on Intrusion Detection System (IDS) and Internal Intrusion Detection and protection system (IIDPS)", 2017International Conference on Inventive Computing and Informatics (ICICI), Pages: 949 – 953

[15]. Yi Yi Aung, MyatMyat Min, "A collaborative intrusion detection based on K-means and projective adaptive resonance theory", 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Pages: 1575 – 1579

[16]. Tahir Mehmood, Helmi B Md Rais, "Machine learning algorithms in context of intrusion detection", 2016 3rd International Conference on Computer and Information Sciences (ICCOINS), Pages: 369 – 373