International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-7, May 2020

Voting Classification Method for Network Traffic Prediction

Altaf Hussain Shah., M. Mazhar Afzal

***Abstract:** The prediction analysis is the approach of data mining which is applied to predict future possibilities based on the current information. The network traffic classification is the major issue of the prediction analysis due to complex dataset. The network traffic techniques have three steps, which are pre-processing, feature extraction and classification. In the phase of pre-processing data set is collected which is processed to removed missing and redundant values. In the second phase, the relationship between attribute and target set is established. In the last phase, the technique of classification is applied for the classification. 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 using voting classifier, which aims at classifying the data into malicious and non-malicious with more accuracy than existing methods.*

***Keywords:** Network traffic analysis, feature extraction, classification, UCI repository*

I. INTRODUCTION

Traffic classification plays an important role in network security and management. A good understanding of applications and protocols in network traffic is necessary for network controller for implementing appropriate security policies. In recent times, Network Traffic Classification is a very important area of Computer Science. The process of identifying the network applications or protocol existing within a network is called network traffic classification [1]. Managing the entire performance of a network is imperative for internet service providers (ISPs). Traffic classification is the primary step in the direction of identifying and classifying unknown network classes. Network traffic classification allows network operators to take some necessary measures like block of some messages and resource management. This approach plays an important role in the growth of network applications. Figure 1.1 shows the general process of network traffic classification.

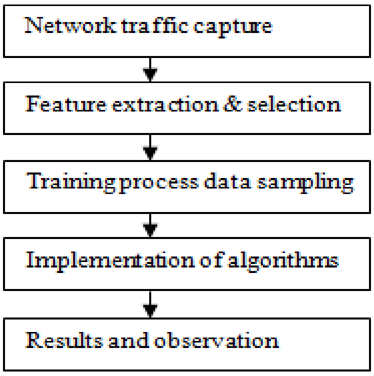
Network traffic capture
Feature extraction & selection
Training process data sampling
Implementation of algorithms
Results and observation

Figure 1.1 General Process of Network Traffic Classification [2]

There are several steps included in the process of network traffic classification. This process depicts the way of using network traffic classification technique for classifying or identifying unidentified network traffic categories. Network Traffic Capturing is the first and very essential step in network traffic classification. This step includes the capturing of the real time network traffic. This step is also called data gathering step. The network traffic can be captured using different available tools. However, Tcpdump tool is the one of the most popular tool used for capturing the real time network traffic. Feature selection and extraction is the next step after capturing the data of network traffic [2]. The extraction of features is done from the captured data in this step. These features include packet period, packet length, inter arrival packet time protocol and so on. Afterward, these features are utilized for the training of machine learning classification model. A tool named Perl script can be utilized for extracting features from captured data. The third step is called training process data sampling. This step performs the sampling of data sets for supervised learning algorithm. Initially, labeling of data is done for the classification of unknown network applications in supervised learning approach [3]. The fourth step involves the implementation of machine learning algorithms. In this step, machine learning algorithm or classifiers are implemented on the examples. There are many tools available for implementing machine learning algorithm. In recent times, most commonly used tools are MatLab and Weka. These tools are known as classification simulation tools. The final step in this process is called result and observation. The simulation tool provides comprehensive results regarding the implemented algorithms in term of different performance metrics after the implementation of machine learning algorithms. These metrics generally include classification accuracy, training time and recall etc. Over the years,

Revised Manuscript Received on May 30, 2020.

***** Correspondence Author

Altaf Hussain Shah.*, Department of computer engineering Glocal University Saharanpur Email: shahaltaf37@gmail.com

Dr M. Mazhar Afzal, Associate Professor & HOD of computer engineering Glocal University Saharanpur, UP. Email: mazhar@theglocaluniversity.in

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

various researchers have proposed different techniques for the classification of network applications [4]. There are mainly three types of network traffic classification techniques. These techniques are known as Port-based Technique, Payload Based Technique and Machine Learning (ML) techniques. Network traffic classification depends on the port numbers in Port-based technique. These port numbers are distributed by the Internet Assigned Numbers Authority (IANA). On a local host in the network, a lot of applications use these allotted port numbers in this technique. The local host is utilized as a meeting point. The other hosts can communicate with each other using this local host. It is imperative for the classifier occurring within the network to verify TCP SYN (TCP SYNCHRONIZE) packets. This is the first step in TCP's three-way handshake protocol in session beginning [5]. This step is implemented for identifying the server end of a novel client-server TCP connection. The target port number of TCP SYN (TCP SYNCHRONIZE) packet is used to direct the application towards a particular port. The second types of techniques are called payload-based techniques [6]. These techniques are also identified as Deep Packet Inspection (DPI). In these techniques, the quality signatures of network applications are submitted in the traffic for observing the packets' content. Almost all payload based approaches analyze the contents of the packet. These techniques make attempt to match these signatures with various other signatures stored in the database. In contrast to port based approaches, these approaches provide more accurate classification results [7]. These techniques are generally utilized for the classification of P2P traffic. One of the most popular ways of network traffic classification is machine learning. Different machine learning algorithms implemented for Traffic Classification are based on supervised and unsupervised learning. However, few approaches may use hybrid algorithms as well. An inclusive labelled data set is required by the supervised learning approach or classification technique. The main objective of this approach is to determine a model or function describing the data. Afterward, this approach is used for the classification of unknown classes [8]. On the other hand, unsupervised machine learning technique approach focuses on determining patterns, structures, or information in non-labeled data.

II. LITRATURE REVIEW

Hyun-Kyo Lim, et.al (2019) used deep learning models for classifying network traffic [9]. For this purpose, some datasets based on packet were created through the pre-processing of network traffic. In order to classify network traffic, convolutional neural network (CNN) and residual network (ResNet) were used. These approaches were used for the training of five deep learning models. At last, the f1 score of the CNN and ResNet deep learning models was used for analyzing the performance of packet-based datasets in the classification of network traffic. The obtained outcomes depicted the efficiency of deep learning models in network traffic classification.

Fakhroddin Noorbehbahani, et.al (2018) recommended a novel semi-supervised algorithm for classifying network traffic [10]. The proposed approach was based on x-means clustering algorithm and a novel label propagation method. A dataset called Moore was used in this work for testing the classification accuracy of recommended approach. The recommended approach provided classification accuracy of 0.95. This result demonstrated the efficiency of this approach in network traffic classifier learning with partial labeled data. A semi-supervised feature selection technique would be used in nearby future for making improvement in the recommended approach.

Xunzhang Li, et.al (2018) proposed a novel pattern matching real-time traffic classification technique. The proposed approach was called PM [11]. This approach initially used jpcap for receiving network traffic data in real-time. Afterward, this approach used pattern matching for matching the real-time network traffic features. This phenomenon provided traffic classification. Moreover, the implementation efficacy of the program had been improved by the use of distributed message system called kafka and the parallel computing structure called Spark in significant manner. The tested outcomes depicted that the proposed approach performed well in terms of accuracy.

R. Archanaa, et.al (2017) analyzed various supervised algorithms to deal with the issue related to the classification of internet traffic [12]. A feature set containing 266 features was used for selection of feature subset. From the given subset, merely eight features were selected. The computation time consumed in the training of framework had been reduced by the feature selection technique. This phenomenon provided improved results as well. The results of the supervised machine learning approaches had been provided in terms of different performance metrics. The result revealed that DECORATE algorithm from the family of Ensemble classification algorithms showed better classification accuracy of 99.65% than all other existing supervised machine learning algorithms.

Xingchao Bian, et.al (2018) used PSO technology for making improvements in the existing network classification algorithms [13]. In this work, the learning approach of semi-supervision had been introduced as well. This approach solved the issues related to time and efficiency. These issues could not be ignored in supervised learning. Moreover, a network traffic classification framework had been designed to enhance the existing algorithm. In contrast to KNN algorithm, PSO optimized KNN algorithm showed more convergence speed as per achieved tested results.

Yaojun Ding, et.al (2016) recommended an ensemble feature selection based classification technique [14]. The proposed approach improved the classification efficiency of large scale imbalanced network traffic. In this work, the frequency matrix had been created on the basis of supervised machine learning algorithm in several data sets. Dataset support had been described for determining the best feature set. The dataset given by Moore of Cambridge was used in this work for testing the efficiency of ensemble feature selection approach. The tested results proved the supremacy of novel algorithm over general supervised machine learning algorithm.

Zhengwu Yuan, et.al (2016) stated that C4.5 decision tree was a very popular supervised classifier [15]. This classifier was mainly used for classifying network traffic. However, increase in data volume reduced the efficacy of this classifier. Hadoop platform was an open source cloud system. This approach showed good performance with big data. Therefore, this approach was mainly used to handle large data.

In contrast to original C4.5 algorithms, the modified algorithm was simpler. The proposed algorithm was similar to the Hadoop platform. This algorithm was named as HAC4.5 decision tree algorithm. The tested outcomes revealed that the modified algorithm along with improving running speed improved the computation accuracy as well.

Wenlong Ke, et.al (2017) proposed a superfluous window-based best feature subset discovery approach for feature selection [16]. The growth algorithm had been used to perform feature selection. Discovering relevant features was the main aim of proposed algorithm. In this work, shrink algorithm had been used for eliminating the redundant features. The integration of Window redundancy and a parallel computing system namely Spark had been done in the algorithm. This phenomenon improved the effectiveness of the algorithm in significant manner. The tested outcomes depicted that the proposed approach performed well in terms of accuracy and scalability. The proposed approach improved the implementation effectiveness of feature selection and traffic classification.

Author	Year	Description	Result
Hyun-Kyo Lim, Ju-Bong Kim, Joo-Seong Heo, Kwihoon Kim, Yong-Geun Hong, Youn-Hee Han	2019	Used deep learning models for classifying network traffic. For this purpose, some datasets based on packet were created through the pre-processing of network traffic	The obtained outcomes depicted the efficiency of deep learning models in network traffic classification
Fakhroddin Noorbehbahani, Sadeq Mansoori	2018	Recommended a novel semi-supervised algorithm for classifying network traffic. The proposed approach was based on x-means clustering algorithm and a novel label propagation method	This result demonstrated the efficiency of this approach in network traffic classifier learning with partial labeled data
Xunzhang Li, Yong Wang, Wenlong Ke, Hao Feng	2018	Proposed a novel pattern matching real-time traffic classification technique. The proposed approach was called PM	The tested outcomes depicted that the proposed approach performed well in terms of accuracy.
R. Archanaa, V. Athulya, T. Rajasundari, M. Vamsee Krishna Kiran	2017	Analyzed various supervised algorithms to deal with the issue related to the classification of internet traffic. A feature set containing 266 features was used for selection of feature subset	The result revealed that DECORATE algorithm from the family of Ensemble classification algorithms showed better classification accuracy of 99.65% than all other existing supervised machine learning algorithms
Xingchao Bian	2018	Used PSO technology for making improvements in the existing network classification algorithms. In this work, the learning approach of semi-supervision had been introduced as well	In contrast to KNN algorithm, PSO optimized KNN algorithm showed more convergence speed as per achieved tested results.
Yaojun Ding	2016	Recommended an ensemble feature selection based classification technique. The proposed approach improved the classification efficiency of large scale imbalanced network traffic	The tested results proved the supremacy of novel algorithm over general supervised machine learning algorithm.
Zhengwu Yuan, Chaozheng Wang	2016	Stated that C4.5 decision tree was a very popular supervised classifier [15]. This classifier was mainly used for classifying network traffic	The tested outcomes revealed that the modified algorithm along with improving running speed improved the computation accuracy as well.

Wenlong Ke, Yong Wang, Xiaochun Lei ,Bizhong Wei	2017	Proposed a superfluous window-based best feature subset discovery approach for feature selection [16]. The growth algorithm had been used to perform feature selection	The proposed approach improved the implementation effectiveness of feature selection and traffic classification.

III. RESEARCH METHDOLOGY

The network traffic 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, k-mean clustering method is applied that clusters the data against being similar and dissimilar. To refine the dataset given in the form of input, the redundancy and missing values are few problems that are eliminated. To calculate the central point of network, k-means clustering approach is applied in the second step. The arithmetic mean of complete dataset is calculated here. The Euclidian distance is calculated from the central point such that the similar and dissimilar points are distinguished. Similar data points are included by one cluster. The data points in separate clusters are dissimilar. SVM classifier is applied in the last step such that the data points can be categorized among two separate classes. To improve the accuracy and performance of classification method, KNN classifier is used that clusters the un-clustered points. It also calculates the Euclidian distance and separates the similar and dissimilar kind of data.

Support Vector Machine

For performing text categorization, a popular predictive model known as SVM classifier is applied. Through data classification, an input is given and the output is achieved by categorizing the data into two categories. For text corpus, the SVM training algorithm is implemented in the mode. Here, in any of the two classes the training sample belongs. An N-Dimensional hyperplane is constructed to categorize the data. To separate the data, two parallel hyper planes are generated on each side of hyper plane. Here, to separate the data such that the distance among two hyper planes can be increased, the hyper plane is used. In correspondence to the division of hyper plane f(X) a linearly separable data set is used for which a linear classification function can be generated. This hyperplane crosses the middle of two classes are separate them from each other. Following is the classification of a new data instance Xn which is achieved by testing the sign of function f(Xn) after its determination:
Where Xn belongs to a positive class if f(Xn) > 0
For larger margin or distance, generalizing the error of classifier is possible. For high dimensional feature set, the performance of an algorithm is known to be better. This algorithm also applied the kernel method such that a new linearly separable data can be achieved by transforming the non-linearly separable data. To calculate regression analysis and perform several calculations, SVM is applied. Further, the elements are ranked by this algorithm. Even when only specific cases are accessed to train, the performance of SVM is better for several attributes. However, the training and testing phases of this algorithm face issues like speed and size. Selecting the kernel function parameters is not easy in this algorithm which is another issue.

K nearest Neighbor

The simplest classification algorithm applied in several applications is called the KNN classifier. Since there are no assumptions included in the underlying data distribution, it is also called the non-parametric supervised learning algorithm. Based on the nearest training samples included in the feature space, the samples are classified. Along with the labels of training images, feature vectors are stored that are included in the training process. During the labeling of K-nearest neighbors, the unlabelled question point is ruled out during classification. Based on these labels, the majority share characterized the object. The object is classified as the class of object that is nearest to be even at k=1. K is considered as an odd integer in case when only two classes are present. To perform multiclass categorization, a tie in the values can also occur in case when k is an odd whole number. Classifying the samples in the basis of majority class of its nearest neighbor is the most important task of KNN algorithms.

$$Class = arg_v max \sum_{(X_i,y_i)\in D_z} I(v = y_i) \qquad \ldots (1)$$

In the above equation, the class label is denoted as v and for ith nearest neighbors, the class label is denoted by $\mathrm{y_i}$. The indicator function is denoted by I. here, if the argument is true, 1 is returned or else 0. In the class of its k-nearest neighbors, the samples are assigned. Initially, to calculate the distance among objects, a distance or similar metric is calculated. Secondly, the set of labeled objects is identified. The total numbers of nearest neighbors in which k value is included is also a parameter that is considered as important. To make the recognition task a success, selecting an appropriate similarity function and value for k parameter is important. Performing KNN classification is very easy and also the classification techniques can be implemented easily.

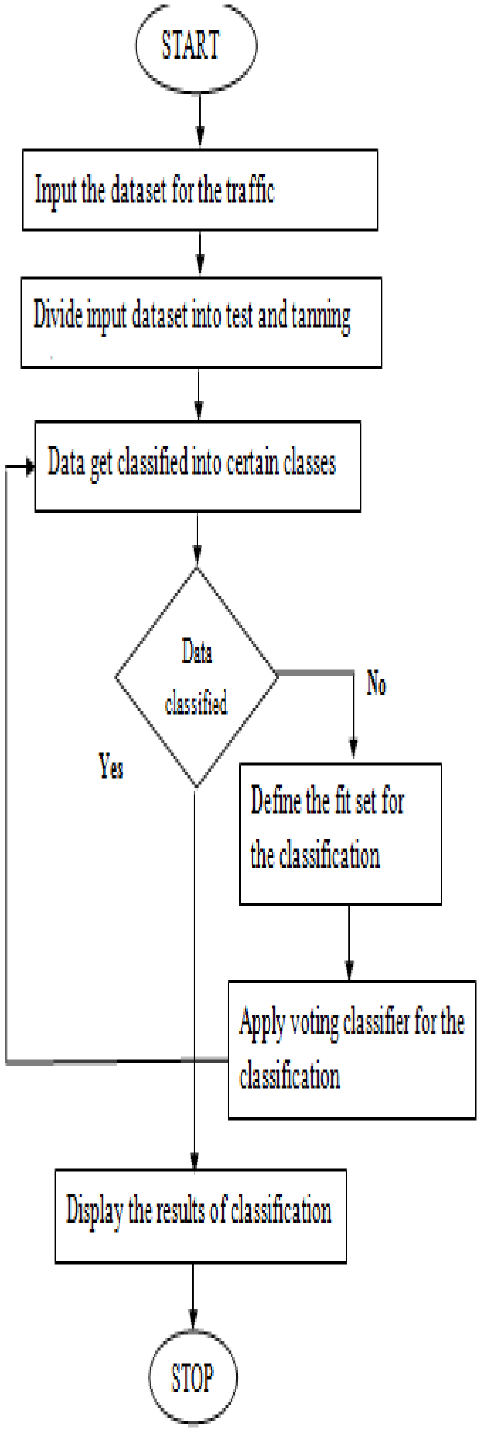
START
Input the dataset for the traffic
Divide input dataset into test and tanning
Data get classified into certain classes
Data classified
No
Yes
Define the fit set for the classification
Apply voting classifier for the classification
Display the results of classification
STOP

Fig 1: Proposed Methodology

IV. RESULT AND DISCUSSION

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.

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

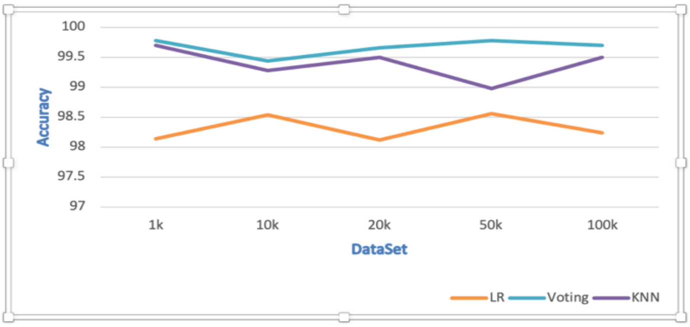

Fig 2: Accuracy Comparison in training phase

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.

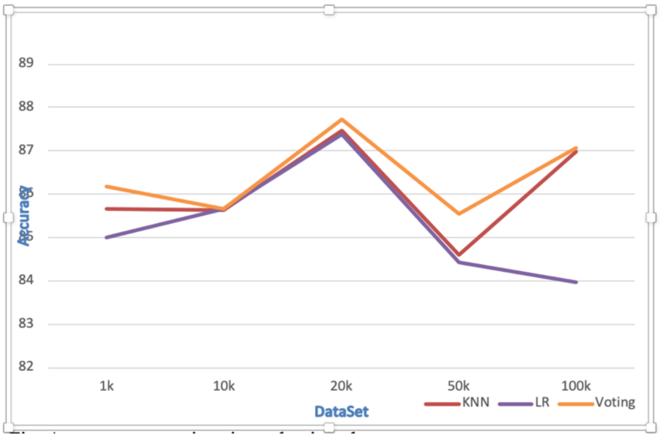

Fig 3: Accuracy Comparison in training phase

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, 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.

REFRENCES

1. Aafa J S, Soja Salim, "A Survey on Network Traffic Classification Techniques", International Journal of Engineering Research & Technology (IJERT), Vol. 3 Issue 3, March – 2014
2. Muhammad Shafiq, Xiangzhan Yu, Asif Ali Laghari, Lu Yao, Nabin Kumar Karn, Foudil Abdessamia, "Network Traffic Classification techniques and comparative analysis using Machine Learning algorithms", 2nd IEEE International Conference on Computer and Communications (ICCC), Year: 2016 | Conference Paper | Publisher: IEEE
3. Dr R Suguna, Suriya Prakash J," A SURVEY ON NETWORK TRAFFIC CLASSIFICATION TECHNIQUES", International Journal of Pure and Applied Mathematics, Volume 117 No. 22 2017, 107-111
4. Supriya Katal, Asstt. Prof. Hardeep Singh," A Survey of Machine Learning Algorithm in Network Traffic Classification", International Journal of Computer Trends and Technology (IJCTT) – volume 9 number 6– Mar 2014
5. Pooja MEHTA, Ruchil SHAH, "A Survey of Network Based Traffic Classification Methods", Database Systems Journal vol.VII, no.4/2016
6. Yoga Durgadevi Goli, R Ambika, "Network Traffic Classification Techniques-A Review", International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), Year: 2018 | Conference Paper | Publisher: IEEE
7. Jun Zhang, Xiao Chen, Yang Xiang, Wanlei Zhou, Jie Wu, "Robust Network Traffic Classification", IEEE/ACM Transactions on Networking, Year: 2015 | Volume: 23, Issue: 4 | Journal Article | Publisher: IEEE
8. Shan Suthaharan, Laxmi Sunkara, Sweta Keshapagu, "Lame' curve-based signature discovery learning technique for network traffic classification", IEEE International Conference on Intelligence and Security Informatics, Year: 2013 | Conference Paper | Publisher: IEEE
9. Hyun-Kyo Lim, Ju-Bong Kim, Joo-Seong Heo, Kwihoon Kim, Yong-Geun Hong, Youn-Hee Han, "Packet-based Network Traffic Classification Using Deep Learning", International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Year: 2019 | Conference Paper | Publisher: IEEE
10. Fakhroddin Noorbehbahani, Sadeq Mansoori, "A New Semi-Supervised Method for Network Traffic Classification Based on X-Means Clustering and Label Propagation", 8th International Conference on Computer and Knowledge Engineering (ICCKE), Year: 2018 | Conference Paper | Publisher: IEEE
11. Xunzhang Li, Yong Wang, Wenlong Ke, Hao Feng, "Real-Time Network Traffic Classification Based on CDH Pattern Matching", 14th International Conference on Computational Intelligence and Security (CIS), Year: 2018 | Conference Paper | Publisher: IEEE
12. R. Archanaa, V. Athulya, T. Rajasundari, M. Vamsee Krishna Kiran, "A comparative performance analysis on network traffic classification using supervised learning algorithms', 4th International Conference on Advanced Computing and Communication Systems (ICACCS), Year: 2017 | Conference Paper | Publisher: IEEE
13. Xingchao Bian, "PSO Optimized Semi-Supervised Network Traffic Classification Strategy", International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), Year: 2018 | Conference Paper | Publisher: IEEE
14. Yaojun Ding, "Imbalanced network traffic classification based on ensemble feature selection", IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), Year: 2016 | Conference Paper | Publisher: IEEE
15. Zhengwu Yuan, Chaozheng Wang, "An improved network traffic classification algorithm based on Hadoop decision tree", IEEE International Conference of Online Analysis and Computing Science (ICOACS), Year: 2016 | Conference Paper | Publisher: IEEE
16. Wenlong Ke, Yong Wang, Xiaochun Lei ,Bizhong Wei, "Spark-Based Feature Selection Algorithm of Network Traffic Classification", 13th International Conference on Computational Intelligence and Security (CIS), Year: 2017 | Conference Paper | Publisher: IEEE.

AUHORS PROFILE

Mr Altaf Hussain shah Received his Bachelor's Degrees from University of Kashmir Srinagar J&K, India, followed by Masters in Networking Administration from Jetking Bangalore India, and Master of Computer Applications from BGSB University Rajouri J&K. Mr Altaf Hussain Shah is a research scholar at Glocal University Saharanpur UP India, his research Interests include Artificial Intelligence and Machine Learning, Network Security, Cloud Security. Altaf Hussain Shah recently Joined US based IT company as NOC Engineer.

Dr M. Mazhar Afzal is an associate Professor and Head of Computer Science Department at Glocal University Saharanpur UP India where he has been since 2015.From 2018 to 2013 he served in a Government University at KSA. He received his Doctorate from Babasaheb Ambedkar Marathawada University Aurangabad (MS). His research span both internet Governance and Network security. Much of his work has been on improving the understanding, design and Performance of Security Systems and various cryptographic techniques In the Networking field he has worked on characterizing the internet and the World Wide Web. In addition, he has also served various academic bodies at different capacities additionally he is also working as Director (IQAC) at Glocal University.