JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Hybrid Classification Model for Network Traffic Classification using SMOTE and PCA

[1]Namita, [2]Rahul

[1]Research Scholar, [2]Assistance Professor
[1,2]Department of Computer Science and Engineering,
Ganga Institute of Technology and Management, Kablana,

Abstract
Network Traffic classification (NTC) involves identifying diverse kinds of applications or traffic data for which the received data packets are analyzed. It is an important technique in communication networks, in recent times. The process to classify the network traffic classification is executed such as to take the data for input, pre-process the data, extract the attributes, classification, and performance analysis. Classifying the network traffic is a crucial process to manage the network and rapid advances in machine learning have driven the utility of learning methods to categorize traffic across a network. This work introduces a new framework to classify the traffic in network so that the efficacy is enhanced. The proposed model is the combination of SMOTE which solve problem of class unbalancing, PCA for the feature reduction and for the classification three classifiers are merged together which are SVM, KNN and random forest. The performance of proposed model is tested in Python is executed using Anaconda to test the efficacy of the introduced framework.

Keywords
Network Traffic Classification, SMOTE, PCA, SVM, KNN, Random Forest

1. INTRODUCTION

The concern related to privacy and security is occurred due to the rapid advancement of Internet. Therefore, the extensively utilized methods to preserve privacy are developed to deal with these issues which are known as VPNs and AMs. The proxy sites are significant for the web surfers these sites assist them in obscuring the quality of the content whose transmission is done to share information as well as in any hidden object [1]. To classify the traffic become a significant field to implement QoS (quality-of-service), generate the traffic and protect the network. Generally, it defines the nature of traffic which a network object has developed. The technique to classify the traffic plays a considerable role in detecting various attacks. Diverse kinds of systems and applications are developed which emphasize on the potential and control of network. The researchers make the deployment of diverse classification with the help of ML (Machine Learning) methods. The resources of network are allocated, controlled and managed using these methods. The classification is the major part of IDS (intrusion detection systems) for detecting the malicious activities. The research community focuses on classifying the network traffic and it becomes major necessity in the industrial field [2]. Several techniques have been presented and constructed over last decades. Diverse kinds of techniques utilized to classify the network traffic are discussed as:

a. Port-based classification: This approach is implemented to recognize an application for which packet header are examined and matched with the TCP or UDP port number whose registration is done with IANA (Internet Assigned Numbers Authority). First of all, some modem applications are flown through the unregistered port numbers or picked a random port, for instance, P2P applications that lead to maximize the FN (false negative). However, other applications focused on concealed themselves for avoiding their filtering due to which the restrictions of operation system access control are avoided.

b. Payload-based classification: Various industry products and works are suggested for dealing with the limitations and reliance of earlier techniques on the basis of conducting investigation ahead of the headers of the packets to contents. This method is known as payload-based classification [3]. The packet contents are discovered and matched with a deterministic set of stored signatures to develop this approach. This approach often provides optimal results with higher accuracy. In addition, this approach is considered as an initial phase in Intrusion Detection System for recognizing the malicious activity in the network.

c. Statistical classification: This approach is designed on the basis of a rationale-based method in which statistical attributes of network traffic flow are deployed for recognizing the application. A number of flow-level metrics such as packet existence, packet IAT, packet lengths and traffic flow idle time are evaluated in this classification. These metrics are unique and effective for particular kind of applications. This assists the classification algorithm in differentiating diverse applications from one another.

d. Behavioral classification: The kind of classification approach emphasizes on observing the entire network traffic which is generated through the endpoint. The network traffic patterns whose generation is done from the target host are analyzed for recognizing the kind of application. To illustrate, the computation of number of communicated hosts is done on the basis of transport layer protocol and the number of ports. This approach is capable of attaining optimal outcomes at least computational cost.

The architecture of NTC model is represented in figure 1. Various phases are comprised in this architecture, in which data is gathered; attributes are extracted; attributes are diminished and selected and a model is constructed [4]. This procedure represents the way of NTC techniques for recognizing and classifying the unknown forms of network traffic with the implementation of ML algorithms.

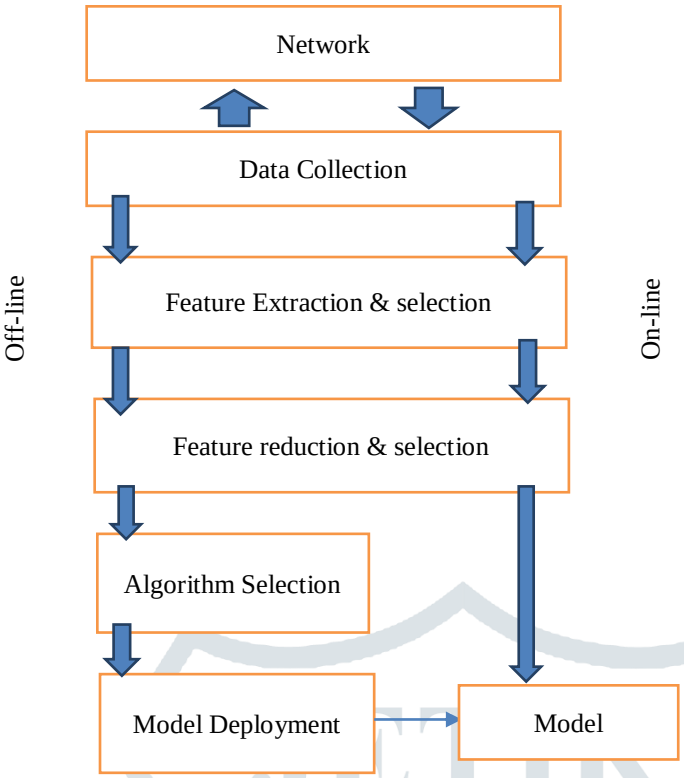


Figure 1: Network Traffic Classification Model

All the phases which are included in the model of classifying the network traffic are explained as:

a. Data Collection: The historical data is considered as a knowledge base in order to develop the ML (machine learning) algorithms. The potential and generality of these algorithms is maximized using different and comprehensive set in which the information regarding an issue is included. Though, this factor is efficient to classify the network traffic because of various qualities such as the web networks are complex and scalable, traffic is developed rapidly and privacy rules do not facilitate to collect the data. In the phase of collecting data, diverse situations are evaluated over the network. This process is focused on collecting the IP runs within a timeframe [5]. In addition, this phase has various tasks, in which packet is managed and flow is reconstructed and stored. The collection of historical dataset is required in the offline flow. However, the online run focuses on the flow of packets.

b. Feature extraction: The major purpose of this phase is to extract the attributes and record the data later on through which issue is put forward. This phase is significant for quantifying the attributes in which information related to the process status is comprised [6]. Moreover, the technique of extracting the attributes is utilized to compute several parameters that represent the significant attributes in the collected data. The main purpose was to acquire the descriptors. The output of this procedure is obtained as a structured table. The feature columns lead to produce this table. A label is a current position of every sample, represented via each row. This row is present in the form of pattern having an extra random column. No label is assigned to the patterns in case of unknown status.

c. Feature selection and reduction: In this phase, the techniques of selecting and reducing the attributes are implemented for acquiring lower space or a set of novel attributes. Various extracted attributes are chosen and reduced using this process. The attributes are reduced to generate novel attributes with the help of original features and the features are selected to discover a reduced set in which a process is described. Various issues are tackled in this phase such as time usage and cost etc. Filters, Wrappers and Embedded techniques are diverse categories of these techniques. ML algorithms are deployed to devise them [7].

d. Classification: The selected features are considered to create a new dataset from the original dataset. The new dataset is adopted in the offline run to generate the models in order to classify the data. The block of selecting an algorithm had processes and methods to select the suitable ML model. This approach has potential to investigate several solutions using various ML algorithms. The identification of an appropriate model is required for various ML techniques so that the traffic can be classified.

2. LITERATURE SURVEY

Hassan Alizadeh, et al (2020) introduced an original GMMs (Gaussian Mixture Models) based scheme to classify and validate the traffic across network on time [11]. The models developed were applied to classify traffic as well as its validation. Classification of traffic was aimed at identifying the source application which was the origin of traffic, based on the evaluation of the highest posterior probability, while traffic validation was focussed on verifying the application claiming to be the traffic originator through probability testing. The new scheme of classifying traffic performed better other existent machine learning based schemes. This scheme obtained classification accuracy of 97.7%with the first 9 packets of a stream. It was also noticed that training the GMM with just 0.5% of overall streams could provide classification accuracy of 96.6%.

Won-JuEom, et.al (2021) used SDN (software-defined networking) framework to develop a structure for classifying traffic across network [12]. The new structure was completely deployed in the network controller to take advantage of the SDN controller's superior computational capability, universal visibility and programmability to classify traffic practically, adaptively and accurately. Several universal metrics such as recall, classification time, training time, accuracy, F1-score were considered in this work to examine the classification performance of four applied ensemble algorithmic approaches. The results of tests demonstrated thar

traffic classification frameworks based on ensemble learning outclassed the other existent framework modelled on practical traffic datasets and presented architecture. In particular, the performance of Light GBM model was the most optimal in the context of classification.

Madhusoodhana Chari S., et.al (2019) discussed the significance of classifying encrypted network traffic in the current period of time [13]. The key aim was to improve the network manageability and to make traffic more visible. The existing researches in this regard discussed the classification of flowing traffic using statistical features since the high cost and inaccuracy of classic strategies made their implementation nearly impossible. Therefore, a scheme based on extracting a packet length signature was presented for classifying the traffic related to various categories. Thus, this work trained a novel J48 decision tree classification structure using a feature set to identify categories of flown traffic. This work also discussed the framework's intelligibility.

Jing Ran, et.al (2018) devised a framework for classifying network traffic based on 3D-CNN(convolutional neural network) [14]. The devised framework was an effort to implement the fruitful technique of analyzing videos to classify traffic across network by studying the predecessors of deep learning. The 3D CNN had the potential to automate the extraction process of attributes and discover the suitable attributes post repetitions of verification thereby performing classification reasonably. Unlike physically selected attribute, such features provided more information. This work integrated the feature extractor and classification model to ensure the best results due to uncertainty of fruitful results by the most optimal extraction followed by the best classification model in a low collaboration environment. Many tests conducted using the openly available USTC-TFC2016 dataset confirmed the superiority of the presented framework over 1D ConvNet and 2D ConvNet in the context of accuracy. The focus will be on continuously studying the integration of feature extractors and classic benchmark classification models in the times to come.

Jiwon Yang, et.al (2019) introduced an original classification approach based on payload that leveraged unencrypted handshake packets shared between end hosts to determine the safety of the transport layer [15]. This work performed classification using Bayesian Neural Network which used input in the form of information about the cipher set, compression technique, and TLS extension of the handshake packet. A series of tests performed proved the dominance of the introduced methodology over other traditional payload-supported classification frameworks Future work will focus on applying this work to classify the existing secure protocols for more exposure.

Yu Wu, et.al (2018) put forward a proposal to improve the classic TDM-EPON framework as well as the upward transferring efficacy by applying statistical multiplexing in MF systems [16]. There were two methods involved in the improved design: (1) classification of traffic (using machine-learning classifiers to classify upstream traffic as practical and impractical); and (2) idle-data transfer (transferring idle traffic to prevent delivery of redundant EPON frames). The first technique was aimed at exploring two feature-selection techniques with classifier biases to obtain the most optimal classification output. The presented method employed these attributes as input in the next methodology. The experimental results obtained on gathered dataset revealed that the presented technique performed well with regard to accuracy and robustness as compared to others. The proposed algorithm performed exceptionally well concerning traffic load, SNR, by reducing E2E delays below 100 µs.

PratibhaKhandait, et al (2020) defined a competent DPI supported methodology to classify streaming traffic by scanning the payload once [17]. The presented heuristic supported scheme was complex in sublinear search. To test, this work used a data set including streams from 10 varied applications, and proved the efficiency of the new approach in terms of accuracy.The experimentation was implemented in Java with the JnetPcap library. The future work would aim to deploy this concept in C via LibPcap library to achieve better results is the topic for future works.

Guanglu Wei, et.al (2020) presented a CNN based DL architecture for the existing intricate networks for traffic classification [18]. Each network traffic load was transformed into a 2D gray image, which was served as input to the framework. This approach converted network stream classification issues to image classification issues. This work led to the research work of classifying the network traffic flowing by applying Deep Neural Network.This could free research community from substantial feature learning and feature selection tasks with benefits in classification accuracy over conventional methodologies.

3.RESEARCH METHODOLOGY

The chief intend of this work is of classifying the flowing network traffic. To classify the network, various phases are executed such as to pre-process the data, extract the attributes and classify the data. The research methodology is defined as: -

1. Data set input and Pre-processing: - The initial stage is the dataset input in which data gathered from the genuine source named KDD is utilized for input. This study employs a NSL-KDD dataset in which 42 attributes are compromised. The duplicate instances are eliminated to enhance the KDD'99 dataset with the purpose of removing the biased classification results from the dataset. The utilization of only 20% of training data is done. However, various editions of the data set are present. This data is represented in the form of KDDTrain+_20Percent. The technique of SMOTE will be used to solve the problem of class unbalancing.
2. Feature Extraction:- This phase aims to establish the association of every attribute with the target set. The FP (false positive) is a scenario in which a sample seems normal but determined as intrusion. The situation in which a sample is actually an intrusion but determined as normal is known as FN (false negative). In case, the earlier case unable to detect the intrusion that is considered as poor FN. The layered approach makes the utilization of most of the IDSs. This indicates the detection of the intrusion through the other layer in case one is not detecting it. The technique of PCA will be applied for the feature reduction.
3. Classification: - This stage focuses on splitting the entire data into training and testing. This work adopts voting classifier for classifying the network traffic. Multiple classifiers are integrated in this technique with the objective of classifying the network traffic. The predicted set is allocated to the test set in this stage. The multiple classifiers will be merged in this phase and those classifiers will be random forest, SVM and KNN. Random Forest algorithm concentrates on developing a forest of individual DTs (decision trees) which are integrated to generate an ensemble. Every tree considers the majority votes in order to predict the data. The data which is sampled at random is employed with replacement for developing trees using RF (Random Forest) due to the sensitivity of DTs to training data. Additionally, this algorithm is useful for maintaining the feature randomness as every tree is

capable of making its decision on the basis of a random subset of attributes at place of picking each attribute from the dataset. SVM (Support Vector Machine) is supported with two properties in which, kernel is represented and margin is optimized. The SVMs is consisted of the association role as a hyperactive plane which sorts out different classes of information. The least hyper range is investigated into the kernel interplanetary in which entirely working out cases are comprised and the place at which an examination case is located in hyper sphere is decided using this algorithm. KNN (K-Nearest Neighbor) often generates optimal outcomes. In certain domains, this algorithm assists in enhancing the existing algorithms after its integration with prior knowledge. The majority label is utilized among its KNN for classifying unlabelled instance using KNN rule. The efficacy of this algorithm is relies on the distance metric which is employed for identifying the nearest neighbors.

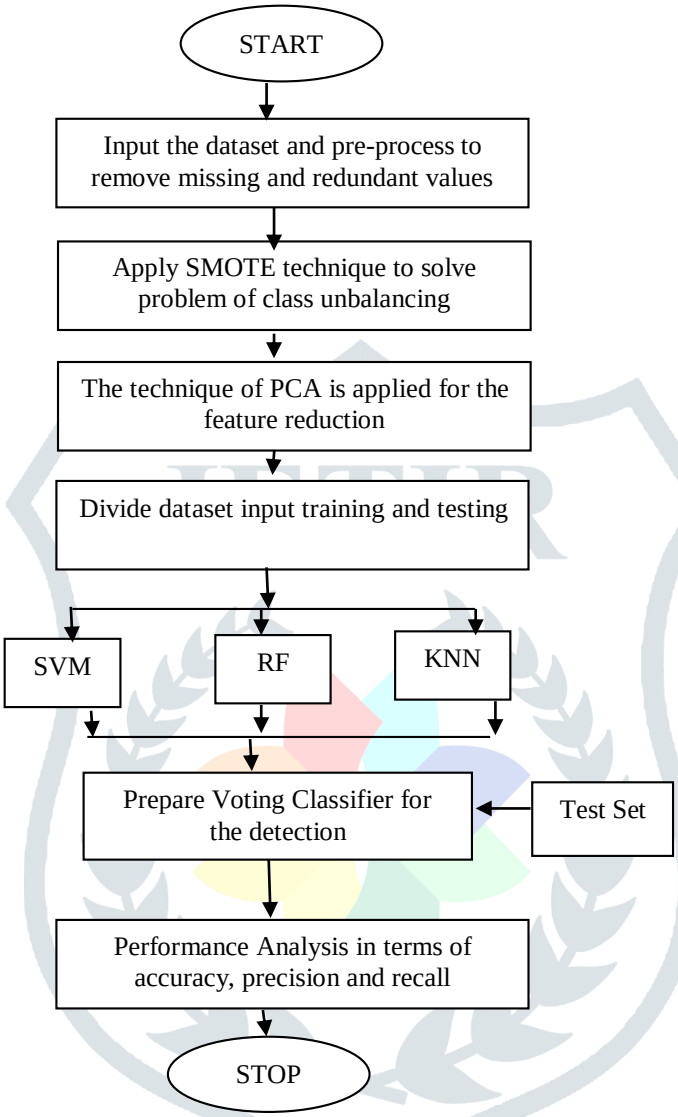


Figure 1: Proposed Methodology

4. RESULT AND DISCUSSION

This research work is conducted on the basis of classifying the network traffic. The framework of classifying the network traffic classification is executed in diverse phases in which the data is pre-processed, features are extracted and the data is classified. The dataset which is used for the model testing is of KDD. The KDD dataset has the 42 attributes and target set which contain multiple classes of different attacks. Various metrics such as accuracy, precision and recall are considered to evaluate the introduced technique.

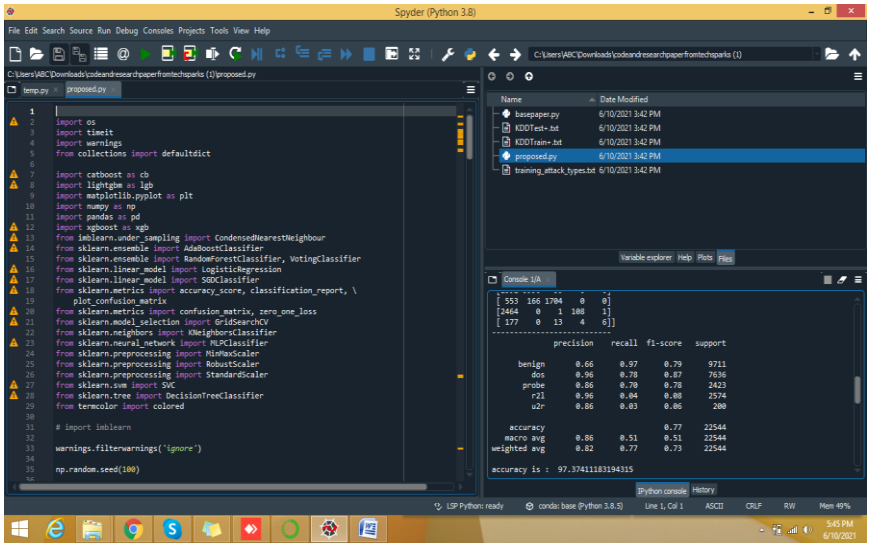


Figure 2: Execution of Proposed Model

As shown in figure 2, the proposed model is executed which is the combination of multiple models. The proposed is the combination of SMOTE, PCA and Voting Classification model for the network traffic classification.

Table 1: Performance Analysis

Model	Accuracy	Precision	Recall
SVM Classifier	75.74 Percent	81 Percent	76 Percent
Logistic Regression	72.67 Percent	80 Percent	77 Percent
KNN Classifier	70 Percent	72 Percent	76 Percent
Random Forest Classifier	75.78 Percent	76.89 Percent	75.90 Percent
Proposed Model	97.37 Percent	82 Percent	77 Percent

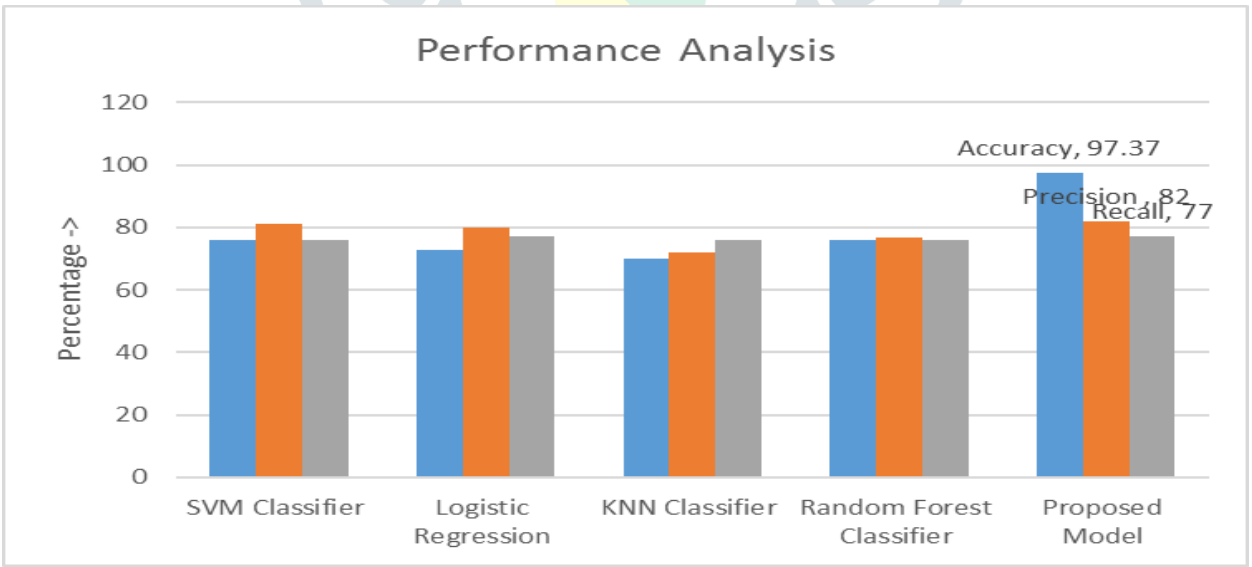


Figure 3: Performance Analysis

Figure 3 depicts the comparison of the introduced technique with various classification models like SVM, Logistic regression, KNN and Random Forest in order to classify the network traffic. The proposed model achieves maximum accuracy of 97.37 percent as compared to other models. The maximum accuracy proves reliability of the proposed model.

CONCLUSION

The network analyst can improve the data with these actions and intrusion detection becomes hard. To automate the process of intrusion detection, effective methodologies are needed which should be very dynamic too. Learn can be performed from those frameworks and the occurrence of any new sorts of intrusions can be detected. Many intrusion detection methods which are highly dynamic, adaptive and perform well with large network traffic have been presented. The novel model is proposed in this research which is combination of multiple techniques and it also solve problem of class unbalancing. In the proposed model SMOTE

technique is used to solve class unbalancing problem. The technique of PCA is used for the feature reduction and in the combination of multiple classifiers are used for the classification. The multiple classifiers are SVM, KNN and random forest. The proposed model achieve accuracy of 97 percent approx. which is approx. 15 percent improvement as compared to other models. In future transform learning techniques can be used for the network traffic classification.

REFERENCES

[1] Jaehwa Park, JunSeong Kim, "A classification of network traffic status for various scale networks", 2013, The International Conference on Information Networking 2013 (ICOIN)
[2] Ji-hye Kim, Sung-Ho Yoon, Myung-Sup Kim, "Study on traffic classification taxonomy for multilateral and hierarchical traffic classification", 2012, 14th Asia-Pacific Network Operations and Management Symposium (APNOMS)
[3] Rui Yang, "The Comparison of Split-Flow Algorithms in Network Traffic Classification: Sequential Mode vs. Parallel Model", 2013, International Conference on Information Technology and Applications
[4] ZebaAtique Shaikh, Dinesh G. Harkut, "A Novel Framework for Network Traffic Classification Using Unknown Flow Detection", 2015, Fifth International Conference on Communication Systems and Network Technologies
[5] ShashikalaTapaswi, Arpit S. Gupta, "Flow-Based P2P Network Traffic Classification Using Machine Learning", 2013, International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery
[6] Sung-Ho Lee, Jun-Sang Park, Sung-Ho Yoon, Myung-Sup Kim, "High performance payload signature-based Internet traffic classification system", 2015, 17th Asia-Pacific Network Operations and Management Symposium (APNOMS)
[7] Yaojun Ding, "Imbalanced network traffic classification based on ensemble feature selection", 2016, IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)
[8] Zhengwu Yuan, Chaozheng Wang, "An improved network traffic classification algorithm based on Hadoop decision tree", 2016, IEEE International Conference of Online Analysis and Computing Science (ICOACS)
[9] Yang Hong, Changcheng Huang, BiswajitNandy, Nabil Seddigh, "Iterative-tuning support vector machine for network traffic classification", 2015, IFIP/IEEE International Symposium on Integrated Network Management (IM)
[10] Chao Wang, Tongge Xu, Xi Qin, "Network Traffic Classification with Improved Random Forest", 2015, 11th International Conference on Computational Intelligence and Security (CIS)
[11] Hassan Alizadeh, Harald Vranken, André Zúquete, Ali Miri, "Timely Classification and Verification of Network Traffic Using Gaussian Mixture Models", 2020, IEEE Access
[12] Won-JuEom, Yeong-Jun Song, Chang-Hoon Park, Jeong-Keun Kim, Geon-Hwan Kim, You-Ze Cho, "Network Traffic Classification Using Ensemble Learning in Software-Defined Networks", 2021, International Conference on Artificial Intelligence in Information and Communication (ICAIIC)
[13] Madhusoodhana Chari S., Srinidhi H., Tamil EsaiSomu, "Network Traffic Classification by Packet Length Signature Extraction", 2019, IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE)
[14] Jing Ran, Yexin Chen, Shulan Li, "THREE-DIMENSIONAL CONVOLUTIONAL NEURAL NETWORK BASED TRAFFIC CLASSIFICATION FOR WIRELESS COMMUNICATIONS", 2018, IEEE Global Conference on Signal and Information Processing (GlobalSIP)
[15] Jiwon Yang, JargalsaikhanNarantuya, Hyuk Lim, "Bayesian Neural Network Based Encrypted Traffic Classification using Initial Handshake Packets", 2019, 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks – Supplemental Volume (DSN-S)
[16] Yu Wu, Massimo Tornatore, Yongli Zhao, Biswanath Mukherjee, "Traffic classification and sifting to improve TDM-EPON fronthaul upstream efficiency", 2018, IEEE/OSA Journal of Optical Communications and Networking
[17] PratibhaKhandait, NeminathHubballi, BodhisatwaMazumdar, "Efficient Keyword Matching for Deep Packet Inspection based Network Traffic Classification", 2020, International Conference on COMmunication Systems &NETworkS (COMSNETS)
[18] Guanglu Wei, "Deep Learning Model under Complex Network and its Application in Traffic Detection and Analysis", 2020, IEEE 2nd International Conference on Civil Aviation Safety and Information Technology (ICCASIT)