STM JOURNALS

Journal of
Artificial Intelligence Research & Advances

ISSN: 2395-6720
Volume 8, Issue 1, 2021
DOI Journal: 10.37591/JoAIRA

http://computers.stmjournals.com/index.php?journal=JoAIRA&page=index

Research JoAIRA

Voting Classification Method for Network Traffic Prediction

Rohan Chauhan*

Abstract

Prediction analysis (PA) is a data mining-based technique. The futuristic outcomes based on present data can be predicted using this technique. As the dataset is large and complex, network traffic classification is a big concern in prediction analysis. Three phases are included in network traffic strategies. The data collection is obtained in the first step of pre-processing, and it is analysed to delete incomplete and obsolete values. The relationship between function and goal set is formed in the second process. In the final step, the classification technique is used to classify the data. The various intrusion attacks on the internet, and also the methods to detect them, also motivated this research work. We reviewed and analysed the well-known network traffic data, NSL KDD dataset and its various features in this research. The proposed model combines Logistic Regression and K-nearest neighbour classifiers with a voting classifier to differentiate data into malicious and non-malicious classes with higher efficiency than current models.

Keywords: Network traffic analysis, feature extraction, classification, UCI repository, KNN classifier.

INTRODUCTION

One cannot deny the significance of TC (Traffic Classification) in the managing of the network. This scheme also makes the network more secure. For network controllers to enforce effective security measures, they must have a clear knowledge of applications and protocols in network traffic. 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. The significance of this approach is increasing with growth of different types of network applications. Figure 1 shows the general process of network traffic classification.

There are several steps that make the classification of network traffic possible. This process depicts the way of using network traffic classification technique for classifying or identifying unidentified network traffic categories. The capturing of traffic within the network is the primary and very essential step in network traffic classification. This step is also called data gathering step. The network traffic can be captured using different available tools. Nonetheless, Tcpdump device is perhaps the most mainstream instrument utilized for catching the ongoing organization traffic. Feature selection and extraction is the next step after capturing the data of network traffic [2].

***Author for Correspondence**
Rohan Chauhan
E-mail: rohan.c199880@gmail.com

Research Scholar, Department of Computer Science & Engineering, University School of Information, Communication and Technology, Gautam Buddha University, Greater Noida 201312, Uttar Pradesh, India

Received Date: January 03, 2021
Accepted Date: March 17, 2021
Published Date: March 26, 2021

Citation: Rohan Chauhan. Voting Classification Method for Network Traffic Prediction. Journal of Artificial Intelligence Research & Advances. 2021; 8(1): 23–30p.

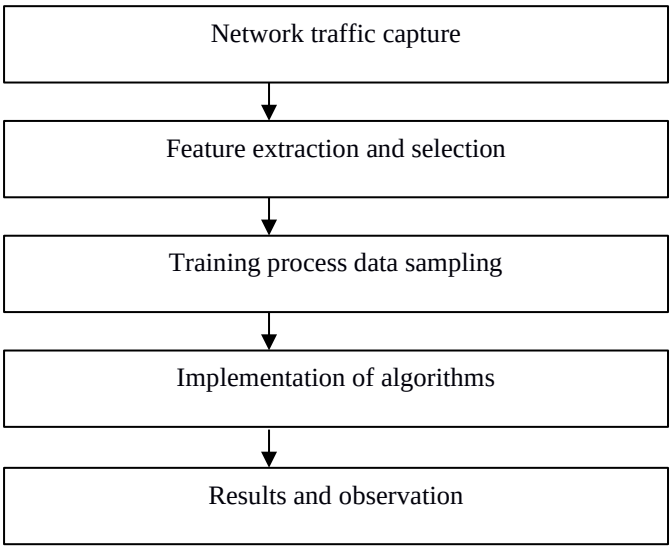

Figure 1. General process of network traffic classification [2].

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. Afterwards, 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 ML (Machine Learning) algorithms. Here, classification algorithms 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 outcomes regarding the implemented algorithms in terms of different performance metrics, once the classification approaches are applied. These metrics generally include classification accuracy, training time and recall etc. Over the years, various researchers have proposed different techniques for the classification of network applications [4]. There are mainly three types of methods based on which the traffic within the network can be classified. These techniques are known as Port-based, Payload-based and ML based 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). A lot of applications use these allotted port numbers in this technique) upon a local network host. The local host is utilized like a meeting spot. 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 [5]. This step is implemented for identifying the server end of a novel client-server TCP link. The objective port number of TCP SYN (TCP SYNCHRONIZE) parcel is utilized to coordinate the application towards a specific 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 functions 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. Rather than port based methodologies, these methodologies give more exact grouping results [7]. These techniques are generally utilized for classifying P2P traffic. A very efficient way of network traffic classification is machine learning. Different machine learning algorithms are implemented to classify traffic. However, some approaches can employ hybrid algorithms as well. An inclusive labeled data set is required by the supervised learning approach or classification technique. The main objective of this approach is to determine a function for describing the data. Afterwards, this approach is applied for the classification of unknown classes [8]. Then again,

unaided AI strategy approach centers around deciding examples, constructions, or data in non-marked information.

LITERATURE REVIEW

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

Noorbehbahani *et al*. recommended a novel semi-supervised approach for classifying the traffic in the network [10]. The proposed approach made use of x-means clustering approach and a novel LP (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.

Li *et al*. 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. Afterwards, 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 using DMS (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.

Archanaa *et al*. analyzed various supervised algorithms to deal with the issues 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. DECORATE algorithm based on Ensemble classification algorithms showed better classification accuracy of 99.65% than all other existing supervised machine learning algorithms as per the outcomes.

Bian *et al*. 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.

Ding *et al*. recommended an ensemble feature selection based classification technique [14]. The proposed approach improved the classification effectiveness of extensive unbalanced network traffic. In this work, the frequency matrix had been created on the basis of supervised machine learning approach in several datasets. Dataset assistance 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.

Yuan *et al*. 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 algorithm, 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.

Ke *et al*. proposed a superfluous window-based best feature subset discovery approach for selecting features [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 a significant manner. The introduced algorithm performed well in terms of accuracy and scalability. The proposed approach improved the implementation effectiveness of FS (Feature Selection) and TC (Traffic Classification) as shown in Table 1.

Table 1. List of literature review.

Authors	Year	Description
Hyun-Kyo Lim, Ju-Bong Kim, Joo-Seong Heo, Kwihoon Kim, Yong-Geun Hong, Youn-Hee Han [9]	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
Fakhroddin Noorbehbahani, Sadeq Mansoori [10]	2018	Recommended a novel semi-supervised algorithm for classifying network traffic. The proposed approach used x-means clustering approach and a novel LP (Label Propagation) method.
Xunzhang Li, Yong Wang, Wenlong Ke, Hao Feng [11]	2018	Proposed a novel pattern matching real-time traffic classification technique. The proposed approach was called PM
R. Archanaa, V. Athulya, T. Rajasundari, M. Vamsee Krishna Kiran [12]	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
Xingchao Bian [13]	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

RESEARCH METHODOLOGY

NTC (Network Traffic Classification) method is performed for classifying 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 and separates them from each other. 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. On the basis of the nearest training instances included in the feature space, the samples are classified. Along with the labels of training pictures, the storing of feature vectors is carried out that are included in the training process. During the labeling of K-nearest neighbors, the unlabelled question point is rejected 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 based on the majority class of its NN (Nearest Neighbor) is the most important task of KNN algorithms.

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

Here, the class label is denoted as v and for ith nearest neighbors, the class label is denoted by 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.

RESULT AND DISCUSSION

NSL-KDD dataset classification method is performed to categorize the traffic against malicious or non-malicious. This technique helps in predicting 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 shown in Figure 2.

The implementation of new research is carried out in Python and the outcomes are evaluated by comparing proposed and existing techniques.

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. Figure 3 demonstrates the results from evaluation phase on test dataset.

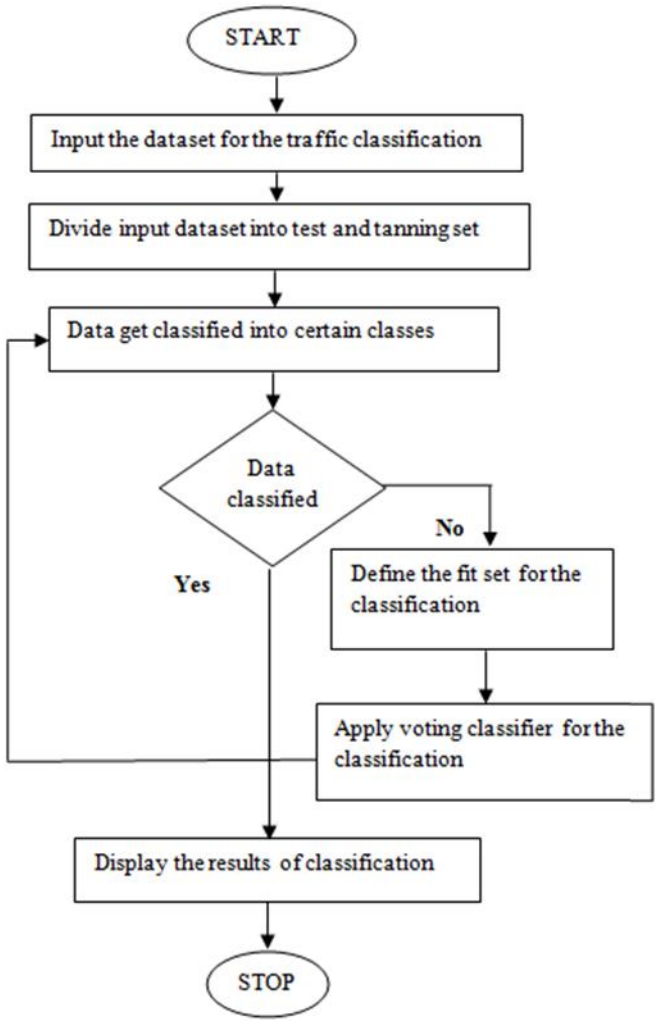

Figure 2. Proposed methodology.

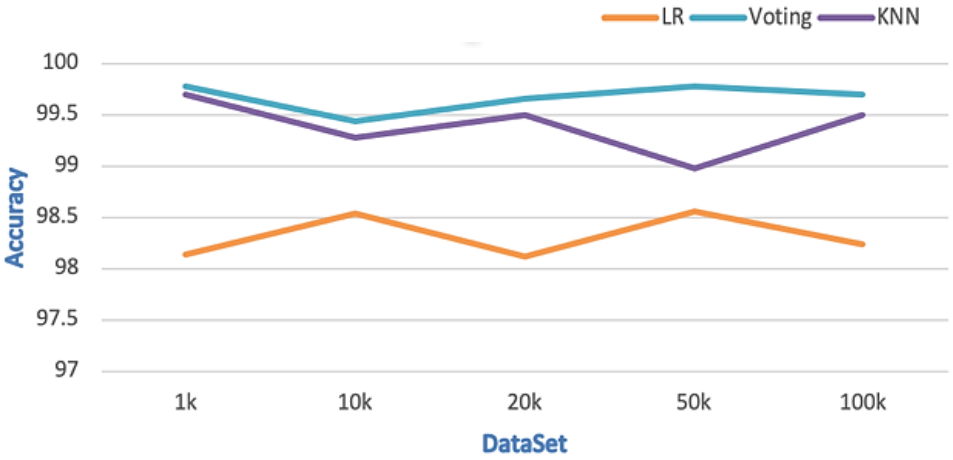

Figure 3. Accuracy comparison in training phase.

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 as shown in Figure 4. But as the dataset increases, Voting starts performing better than KNN and LR which we can see above with 20k, 50k and 100k datasets.

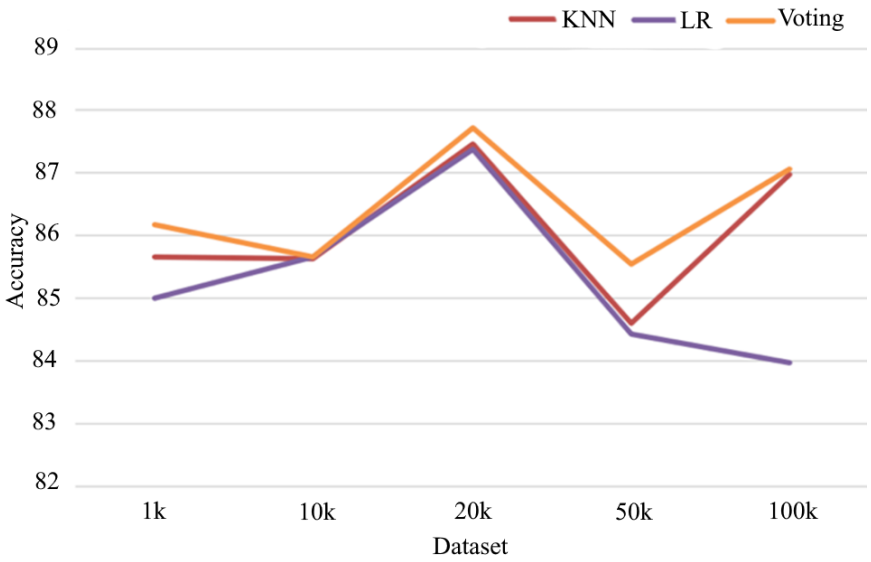

Figure 4. Accuracy comparison in training phase.

CONCLUSION

The research is effective in understanding how the Voting classifier outperforms various classifiers such as LR and KNN. It also highlights the importance of training the platform to optimize the framework's accuracy level. In Voting Classifier, there are a variety of classifiers that could be integrated and used; but, in this study, we used KNN and logistic regression classifiers. The modified classifier shows that the more the training data, the better the algorithm results in the evaluation process.

REFERENCES

1. Aafa JS, Soja Salim. A Survey on Network Traffic Classification Techniques. International Journal of Engineering Research & Technology (IJERT). 2014; 3(3): 2015–2019.
2. Muhammad Shafiq, Xiangzhan Yu, *et al*. Network Traffic Classification techniques and comparative analysis using Machine Learning algorithms. 2nd IEEE International Conference on Computer and Communications (ICCC); 2016 Oct 14–17; Chengdu, China. New York: IEEE; 2017.
3. Suguna R, Suriya Prakash J. A Survey on Network Traffic Classification Techniques. Int J Pure Appl Math. 2017; 117(22): 107–111.
4. Supriya Katal, Hardeep Singh. A Survey of Machine Learning Algorithm in Network Traffic Classification. International Journal of Emerging Trends & Technology in Computer Science (IJETTCS). 2014: 9(6): 301–304.
5. Pooja Mehta, Ruchil Shah. A Survey of Network Based Traffic Classification Methods. Database Syst J. 2017; 7(4): 24–31.
6. Yoga Durgadevi Goli, Ambika R. Network Traffic Classification Techniques-A Review. International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS); 2018 Dec 21–22; Belgaum, India. New York: IEEE; 2019.
7. Jun Zhang, Xiao Chen, *et al*. Robust Network Traffic Classification. IEEE/ACM Trans Netw. 2015; 23(4): 1257–1270.
8. Shan Suthaharan, Laxmi Sunkara, *et al*. Lame' curve-based signature discovery learning technique for network traffic classification. IEEE International Conference on Intelligence and Security Informatics; 2013 Jun 4–7; Seattle, WA, USA. New York: IEEE; 2013.

9. Hyun-Kyo Lim, Ju-Bong Kim, *et al*. Packet-based Network Traffic Classification Using Deep Learning.International Conference on Artificial Intelligence in Information and Communication (ICAIIC); 2019 Feb 11–13; Okinawa, Japan. New York: IEEE; 2019.
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); 2018 Oct 25–26; Mashhad, Iran. New York: IEEE; 2018.
11. Xunzhang Li, Yong Wang, *et al.* Real-Time Network Traffic Classification Based on CDH Pattern Matching. 14th International Conference on Computational Intelligence and Security (CIS); 2018 Nov 16–19; Hangzhou, China. New York: IEEE; 2018.
12. Archanaa R, Athulya V, *et al.* A comparative performance analysis on network traffic classification using supervised learning algorithms. 4th International Conference on Advanced Computing and Communication Systems (ICACCS); 2017 Jan 6–7; Coimbatore, India. New York: IEEE; 2017.
13. Xingchao Bian. PSO Optimized Semi-Supervised Network Traffic Classification Strategy. International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS); 2018 Jan 25–26; Xiamen, China. New York: IEEE; 2018.
14. Yaojun Ding. Imbalanced network traffic classification based on ensemble feature selection. IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC); 2016 Aug 5–8; Hong Kong, China. New York: IEEE; 2016.
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); 2016 May 28–29; Chongqing, China. New York: IEEE; 2016.
16. Wenlong Ke, Yong Wang, *et al.* Spark-Based Feature Selection Algorithm of Network Traffic Classification. 13th International Conference on Computational Intelligence and Security (CIS); 2017 Dec 15–18; Hong Kong, China. New York: IEEE; 2018.

