A Classification Method for Network Traffic Based on Semi-supervised Approach

Yun Liu*, Zhiqiang Zhu and Pengfei Zhong

1 The Second Research Institute of CAAC, Chengdu, Sichuan, 610000, China
*Corresponding author’s e-mail: liyun@cdatc.com

Abstract. Traffic data encryption has been a trend in most Internet applications, and traditional protocol filtering based on fixed port and traffic classification based on Deep Packet Inspection are unable to identify encrypted traffic. Recently, the traffic classification method based on the statistical characteristics of network traffic, which can solve the problem of encrypting data or user privacy protection, has been widely discussed. However, the traditional supervised learning method requires manual marking of a large amount of network traffic data, which is tedious and time-consuming. In this paper, a improved semi-supervised traffic classification framework based on BIRCH clustering method is proposed, and through experiments, the proposed algorithm, supervised learning algorithm and classical semi-supervised traffic classification algorithm are analyzed and compared. The results show that the algorithm proposed in this paper has higher overall accuracy and classification accuracy, and the algorithm can increase the accuracy on traffic classification.

1. Introduction

At present, the traffic management and planning of Internet network becomes more and more difficult. On the one hand, the protocol characteristics of all kinds of new applications become more complex and diversified. On the other hand, various network attacks become too covert to be identified by traditional traffic classification. How to effectively identify and control the new type of network traffic and accurately judge the illegal application traffic is the important work of network traffic management and planning.

Network traffic classification technology can be divided into three categories: protocol filtering technology, deep packet inspection (DPI) technology and traffic classification technology. Protocol filtering based on fixed port can identify the specific port number of different protocols, but it can't do anything for applications using dynamic port number. With checking the application signature in the payload of IP packet, DPI technology can solve the problem of dynamic port, but will completely fail in encrypting data or user privacy protecting. Traffic classification technology based on statistical characteristics of network traffic constructs network flow features vectors, combined with machine learning algorithm, has been widely concerned in the application of network traffic classification.

At the same time, about 80% of the Web traffic in the current Internet network has been encrypted, and the encrypted traffic continues to grow at a rate of over 90% every year [1]. The growing trend of encrypted traffic suggests that the use of unencrypted traffic classification methods will soon become meaningless. The traditional encrypted traffic classification method uses supervised learning to model the data, which requires a lot of tedious and time-consuming manual annotation. This paper designs a semi-supervised traffic classification framework based on BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) clustering, and classifies network traffic by using the improved semi-
supervised clustering method. In experiments, we analyze and compare proposed algorithm, supervised learning algorithm and classical semi-supervised traffic classification algorithm. The results show that the proposed algorithm has higher overall accuracy and classification accuracy, and can accurately classify traffic data of various protocols and can be applied to the current encrypted network data communication.

2. Related work
In the classification method based on network traffic, network flow is defined as having the same quintuple (source IP address, source port, destination IP address, destination port and transport layer protocol) continuous packets, each of the network flow can be represented by one or more network flow characteristics, such as the average packet length and packet to the average time interval. In view of a large number of network flow characteristics, network flows can be transformed into feature vectors, and the classification of network traffic belonging to different protocol types can be realized by combining machine learning technology.

Machine learning algorithms include supervised machine learning, unsupervised machine learning and semi-supervised machine learning.

Supervised machine learning algorithm uses labeled data to train the optimal classification model, and then uses the model to map the unclassified data input to the corresponding output. Classical supervised learning algorithms include support vector machines[2], decision trees[3] and Bayesian networks[4]. Some supervised learning algorithms have been applied to traffic classification [5-6] and show high accuracy. However, this method requires a large amount of labeled data to train the classifier, while manual labeling data is very laborious and costly. In addition, when there is an unknown protocol in the network that does not exist in the training data set, the protocol will be wrongly classified into the protocol category that exists in the training data, which will affect the accuracy of traffic classification of such protocol.

Different from the supervised learning algorithm, the unsupervised learning algorithm does not need the labeled data in training phase. It can divide the unlabeled data with similar characteristics into one cluster, and then finish mapping. The typical algorithms of unsupervised learning algorithm include automatic clustering algorithm[7], k-means algorithm[8], and density clustering algorithm[9], all of which have been applied in the field of traffic classification. The defect of unsupervised learning traffic classification method lies in low accuracy, because the number of clusters must be set large enough, and it is difficult to map a large number of traffic clusters to a small number of traffic categories without the help of labeling information.

Semi-supervised learning is a compromise between supervised learning and unsupervised learning, and its training data is a mixture of labeled and unlabeled data. The traffic classification method based on semi-supervised learning can not only reduce the dependence on labeled data, but also optimize the classifier by utilizing a large amount of easily obtained unlabeled data. Erman et al.[10] first proposed the application of semi-supervised learning method to the field of traffic classification. By using the mixed input clustering of labeled and unlabeled streams, they classify the known traffic and extract the unknown traffic. Then, some improved models based on semi-supervised learning was proposed[11,12], which modify the classification method and increase the classification accuracy and adaptability of complex network environment.

3. Methodology
As discussed earlier, although semi-supervised learning method has great advantages in network traffic classification, there is still room for improvement, especially in the selection of clustering mapping method. This section will discuss in detail the semi-supervised traffic classification framework based on BIRCH clustering.
3.1. Datasets
Considering the privacy and security problems of network traffic and the differences of network traffic distribution in different network environments, we select the ISCX VPN-nonVPN traffic dataset[13] to conduct experiments. After excluding the traffic specially encrypted by VPN, the rest conventional traffic types and the traffic content under this type are shown in Table 1. It contains seven types of network traffic divided by categories of traffic.

| Traffic type | Content               |
|--------------|-----------------------|
| Email        | SMTPS, POP3S, IMAPS   |
| Chat         | ICQ, AIM, Skype, Facebook, Hangouts |
| Streaming    | Vimeo, Youtube, Netflix, Spotify |
| File Transfer| Skype, SFTP, FTPS    |
| VOIP         | Facebook, Skype, Hangouts |
| P2P          | uTorrent, Bittorrent |
| Web Browsing | Firefox, Chrome      |

A total of 14000 network flows of conventional encrypted traffic of 7 categories in the ISCX data set are selected as the experimental data(In order to ensure the fairness of the experimental comparison process, 2000 network streams of each category were selected). The original network flow contains 23 features, among which there is some redundancy. Before the experiment, we need to conduct the following processing on the original network stream.

The first step is to normalize the data. Different network flow features have different dimensions, such as network flow time features with millisecond as the unit, whose value is often in the millions. In the meanwhile, the characteristic value of network flow rate is usually ranges from dozens to hundreds. In order to eliminate the influence of dimensionality, z-score normalization is performed on the data set at first. Then PCA dimensionality reduction is needed for the original network stream. After the normalization of the original network flow data set, the variance sum of the principal components is set to 99% of the variance sum of all the characteristics of the original 23-dimensional network flow features. After calculation, the first 9 principal components can meet the requirements. Therefore, the normalized data is reduced to 9 dimensions in this paper.

3.2. Improved semi-supervised clustering algorithm
3.2.1. Clustering Process
Based on the processing principle of the semi-supervised classification algorithm, we use a small number of labeled samples and a large number of unlabeled samples as the inputs of the BIRCH clustering algorithm. Due to the randomness of the selection of labels with samples, the clustering algorithm must be able to deal with variable clustering density and unbalanced cluster size.

BIRCH(Balanced Iterative Reducing and Clustering using Hierarchies) demonstrates that it is especially suitable for very large database. BIRCH constructs a height-balanced CF-tree with two parameters, branching factor and cluster diameter, to carry out the whole clustering process. The branching factor represents the maximum number of child nodes of each node of the tree, while the cluster diameter is used to determine the sample similarity. If sample similarity is less than the cluster diameter, it belongs to the cluster. Otherwise, it does not belong to the cluster. In following experiment, the cluster size was adjusted mainly by adjusting the cluster diameter, so as to observe the influence of this parameter on the mapping effect of the cluster category.
3.2.2. Improved cluster category mapping

After the clustering phase, a large number of unlabeled samples and a small number of labeled samples are placed in the same cluster. At this point, according to the number of categories of labeled samples in the same cluster, it can be divided into the following three situations:

- Case 1: The same cluster contains only one category of labeled samples;
- Case 2: The same cluster contains different categories of labeled samples;
- Case 3: The clusters did not contain any labeled samples.

For case 1, we map all the categories of the remaining unlabeled samples to this category.

For case 2, the original method counts the categories with more labeled samples in the cluster as the mapping result of the categories of the remaining unlabeled samples. However, when the similarity of network flow samples is not high, the remaining unlabeled samples in the cluster may also belong to different clusters. So, we improve the original method of category mapping and add the criteria for the number of samples labeled by different clusters. By counting the number of each category of labeled samples in the cluster, the maximum number of labeled samples $N_1$ and the secondary number of labeled samples $N_2$ were obtained. In order to map the remaining unlabeled samples in the cluster to the majority of the cluster, the two need to satisfy the following relationship:

$$\alpha = \frac{N_2}{N_1} \quad \alpha < 0.3$$  \hspace{1cm} (1)

Follow-up experiments showed that when the unlabeled samples, generated by the clustering algorithm, do not belong to the same cluster, the method proposed can avoid the wrong labeling of these unlabeled samples to a certain extent.

For case 3, these unlabeled samples may be unknown traffic, and to be examined manually later.

3.2.3. Random forest training

After the process of clustering and improvement of cluster category mapping, a large number of network stream samples with accurate category labeling have been constructed. Input these samples into the machine learning model for model training and a classifier with higher accuracy can be obtained.

Random forest is a combination classification algorithm proposed in 2001[15], which balances the errors of unbalanced samples and has good tolerance to outliers and noise.

We used Bagging method to construct the random forest. In this method, several training subsets are generated randomly, a classifier is trained on each training subset, and the final classification result is generated by voting the classification results of multiple classifiers. The algorithm is stated as follows:

- Step 1: For a given training sample, through $n$ random repeatable samples, construct a bootstrap sample based on the data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$;
- Step 2: Construct a decision tree based on each bootstrap sample;
- Step 3: Repeat step 1 and step 2 to get multiple trees;
- Step 4: Each tree votes on the input vector $x$;
- Step 5: Count all the votes and the one with the highest number of votes was the category label for vector $x$;
- Step 6: The ratio, which is not the same as the correct classification label, is the misclassification rate of random forest.

4. Experiment Results

In order to verify the improved semi-supervised traffic classification algorithm proposed in this paper, a simulation experiment was conducted to compare the proposed algorithm with the classical semi-supervised traffic classification algorithm[10] and with the supervised traffic classification algorithm based on KNN[16]. The evaluation criteria of the experiment were mainly the overall accuracy and classification accuracy, and the experimental results were averaged after 50 replicates.

The system original sample was 7 types of network flows as shown in Table 1, with 2000 records for each type. The network flow samples without category labels are divided into 60% training set and 40% testing set. The training set and a few samples with category labels are used as the input samples of the
semi-supervised classification model, while the testing set is used to simulate the real network flow samples to test the three traffic classification algorithms respectively. The input samples of the supervised classification model are a small number of network stream samples with category labels. In addition, according to the empirical value of the prior experiment, the class diameter of the BIRCH clustering algorithm is selected as 0.5, and other parameters of the two algorithms participating in the comparison are set the same.

The overall accuracy and classification accuracy (F-measure) comparison between three algorithms are shown in Fig. 1 and Fig. 2.

Based on the selected traffic samples, Fig. 1 shows that the overall accuracy of the proposed algorithm is higher than that of the classical supervised and semi-supervised algorithms, indicating that the improvement of the proposed algorithm on clustering and cluster category mapping does improve the overall classification accuracy.

As shown in Fig. 2, the classification accuracy of algorithm proposed for 7 kinds of protocol data is more than 95%. The traditional semi-supervised algorithm and supervised algorithm perform better for Web Browsing, File Transfer and Streaming network streams, while the classification accuracy of other network streams is not good. The experimental results further demonstrate that the algorithm proposed can more accurately implement the classification of multi-protocol network traffic.
5. Conclusion
In this paper, an improved semi-supervised traffic classification framework based on BIRCH clustering method is proposed. A total of 14,000 network traffic samples of 7 kinds were selected as experimental input samples, and the proposed algorithm, supervised learning algorithm and classical semi-supervised traffic classification algorithm were analyzed and compared in the experiment. The results show that the algorithm proposed in this paper has higher overall accuracy and classification accuracy, and the algorithm can increase the accuracy on traffic classification.

References
[1] [EB/OL]. (2019) Encrypted Traffic Analytics Non-Fabric Prescriptive Deployment Guide.https://www.cisco.com/c/dam/en/us/td/docs/solutions/CVD/Campus/eta-nonfab-deploy-guide-2019oct.pdf.
[2] H Kim, KC Caffy, et al. (2008) Internet traffic classification demystified: myths, caveats, and the best practices. In: ACM Conference on Emerging Network Experiment and Technology. Madrid. pp. 9-12,
[3] Williams N, Zander S, et al. (2006) A Preliminary Performance Comparison of Five Machine Learning Algorithms for Practical IP Traffic Flow Classification. Sigcomm Computer Communication Review., 36(5):5-16
[4] Auld T, Moore A W, Gull S F. (2007) Bayesian Neural Networks for Internet Traffic Classification. IEEE Transaction on Neural Networks., 18(1):223-239
[5] Moore A W, Zuev D. (2005) Internet Traffic Classification Using Bayesian Analysis Techniques. ACM Sigmetrics Performance Evaluation Review., 33(1):50-60
[6] Este A, Gringoli F, Salgarelli. (2009) Support Vector Machines for TCP traffic classification[J]. Computer Networks., 53(14):2476-2490
[7] Zander, Sebastian, Nguyen, et al. (2005) Automated Traffic Classification and Application Identification using Machine Learning. Local Computer Networks, the IEEE Conference on Networks., 2005:250-257
[8] Entian J, Mahanti A, Arlitt M. (2006) QRP05-4: Internet Traffic Identification using Machine Learning. In: GLOBECOM. IEEE. pp. 1-6
[9] Entian J, Arlitt M, Mahanti A. (2006) Traffic classification using clustering algorithms. ACM Workshop on Mining Network Data. pp. 281-286
[10] Erman J, Mahanti A, et al. (2007) Offline/realtime traffic classification using semi-supervised learning. Permanace Evaluation., 64(9-12):1194-1213
[11] Zhang J, Chen C, Xiang Y, et al. (2013) An Effective Network Traffic Classification Method with Unknown Flow Detection. IEEE Transactions on Network and Service Management., 10(2):133-147
[12] Zhang J, Chen C, Xiang Y, et al. (2015) Robust Network Traffic Classification. IEEE/ACM Transactions on Networking., 23(4):1257-1270
[13] [EB/OL]. (2021) VPN-nonVPN dataset (ISCXVPN2016) https://www.unb.ca/cic/datasets/vpn.html
[14] Tian Z, Ramakrishnan R, Livny M. (1996) BIRCH: An Efficient Data Clustering Method for Very Large[J]. acm sigmod record., 25(2):103-114
[15] Breiman L. (2001) Random forests. Machine Learning., 45(1):5-32
[16] Wu Di, Chen Xiao, et al. (2014) On addressing the imbalance problem: a correlated KNN approach for network traffic classification. In: Proceedings of International Conference on Network and System Security. Berlin., pp. 138-151