International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 3, ISSN No. 2455-2143, Pages 617-619
Published Online July 2020 in IJEAST (http://www.ijeast.com)

NETWORK CLASSIFICATION AND IDENTIFICATION BASED ON MACHINE LEARNING

Ganga Gudi
Department of Computer Science,
KLE's S.Nijalingappa College,
Bangalore, India

Dr Hanumanthappa M
Department of Computer Science and
Applications, Jnana Bharathi Campus,
Bangalore University, India

Abstract- **The traffic identification is an important basis for traffic monitoring and data analysis. In this paper, the analysis of network traffic identification is based on machine learning and deep packet inspection. This method uses deep packet inspection technology to identify most network traffic, and improves the accuracy of identification. Machine learning method is used to the identify network traffic with encryption and unknown features, which makes up for the disadvantage of deep packet inspection that cannot identify new applications and encrypted traffic.**

Keywords-**machine learning, DPI, network identification.**

I. INTRODUCTION

The rapid development of network technology, where users are demanding higher speed and quality of network services. It has become one of the challenges in the field of network operation and maintenance management to manage and control various network traffic.

There are three commonly used methods:

a) Identification Method based on Port Matching

Identification method classifies the application based on the port number recommended by the IANA (The Internet Assigned numbers Authority). With a large number of network services and development of P2P technology, applications begin to use port and dynamic port technology in order to cross firewalls or avoid other blocking strategies.

b) Identification Method based on Deep Packet Inspection

The DPI (deep packet inspection) technology detects the load content of IP packet during network interaction or data transmission according to the method of pattern matching, and determines the type of application according.

c) Identification based on Machine Learning

Due to different application protocols, network data flow has different characteristics in terms of data flow duration, packet length, packet transmission frequency and packet rate. According to these characteristics of network flow, the identification technology in data mining can be used to achieve good traffic identification through machine learning.

II. NETWORK TRAFFIC IDENTIFICATION METHOD BASED ON MACHINE LEARNING AND DPI TECHNOLOGY

A. Machine Learning and DPI technical Identification Methods

1) DPI technology

DPI technology is based on feature field detection. By deeply reading the IP packet load content and reorganizing the application layer information. Then, the data flow content is scanned and detected according to the existing feature library.

a) Machine learning identification methods

The core of the network traffic identification method based on machine learning is that computer programs can constantly improve their performance with the accumulation of learning experience, so as to complete tasks.

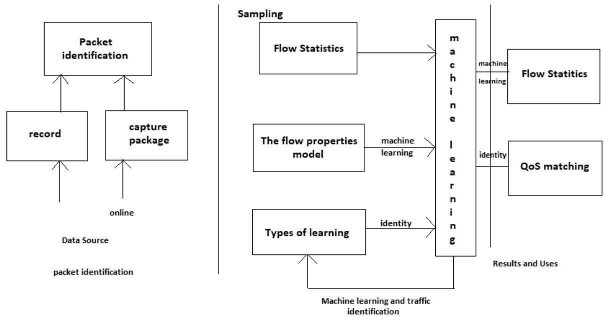

Fig. 1 Flow identification flow chart based on machine learning

B. Algorithm Design

It is the combination of machine learning and DPI technology.

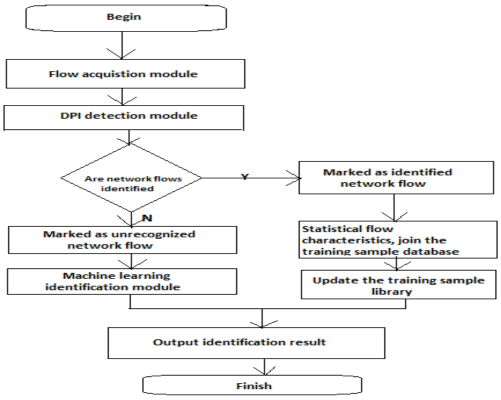
Begin
Flow acquistion module
DPI detection module
Are network flows identified
Y
Marked as identified network flow
N
Marked as unrecognized network flow
Statistical flow characteristics, join the training sample database
Machine learning identification module
Update the training sample library
Output identification result
Finish

Fig. 2 Network traffic identification process based on DPI and machine learning

- In the DPI detection stage, pattern matching detection is carried out for network data flow according to the protocol feature library loaded.
- Once it is identified by DPI, the flow statistical feature acquisition module sets a fixed collection time and begins to collect the feature information. For the network flow identified by DPI, its flow statistical characteristic information is added to the training sample library.

C. Algorithm Implementation

Network traffic identification method is based on DPI and flow statistics features are mainly composed of DPI detection module and machine learning identification module.

a) DPI detection module

The DPI detection module mainly based on the feature library RuleLib, conducts in-depth analysis of data traffic through pattern matching, and identifies specific application traffic. The detection flow table stores the detected data flow according to the quintuple information.

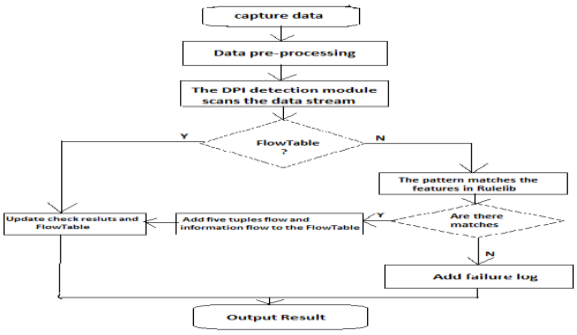
capture data
Data pre-processing
The DPI detection module scans the data stream
Y
FlowTable ?
N
The pattern matches the features in Rulelib
Update check resluts and FlowTable
Add five tuples flow and information flow to the FlowTable
Y
Are there matches
N
Add failure log
Output Result

Fig. 3. DPI network traffic identification process.

b) *Machine learning identification module*

The naive baye's identification method is to determine the category of the sample by calculating the posterior probability. Its basic idea is based on baye's formula and conditional independence assumption in probability theory and the combined probability of attribute and category is used to estimate the category of the new sample.

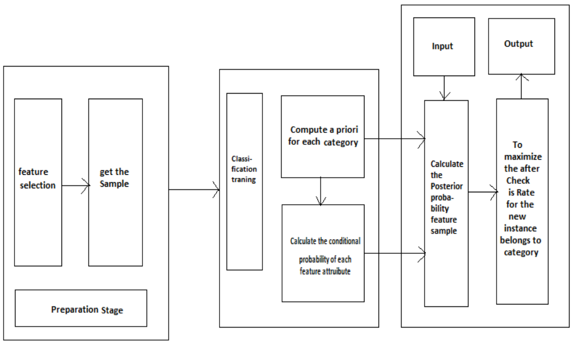
feature selection
get the Sample
Preparation Stage
Classi- fication traning
Compute a priori for each category
Calculate the conditional probability of each feature attruibute
Input
Output
Calculate the Posterior proba- bility feature sample
To maximize the after Check is Rate for the new instance belongs to category

Fig. 4 Naive baye's network traffic identification process

c) *Preparation stage.*

Feature selection is to get the most effective feature combination on the premise of ensuring the original value of the application data. Feature selection adopts FCBF algorithm to select 8 feature attributes from 248 feature attributes of network flow as feature set.

d) *Training stage of classifier*

The naive baye's belongs to the supervised identification method. Matlab is used to analyze the sample data set and calculate the prior probability and conditional probability.

e) *Identification stage.*

The classifier is generated according to the decision model of the training phase to realize the online identification of network traffic.

III. CONCLUSION

By analyzing two kinds of network traffic identification methods based on feature field and flow statistics, a network traffic identification method based on machine learning and DPI technology is proposed. It uses DPI technology to identify most network traffic, reduces the workload. DPI technology can identify specific application traffic, and improves the accuracy of identification. The machine learning method based on the statistical characteristics of flow is used to assist the identification of network flows with encryption and unknown features, which makes up for the shortcomings of DPI technology in identifying new applications and encrypted traffic, and improves the identification rate of network traffic.

IV. REFERENCES

1) Xiaoguang Zhang, Lixia Xi and Congpeng Lu - Research Optical Network Traffic Based on the

Content Identification.

2) Research on Network traffic identification Based on Machine Learning Method, 2009.

3) Jingyu Wang, Jiyuan Zhang and, Yuesheng Tan - Research of P2P Traffic Identification Based on Traffic Characteristics.

4) Li Wei, Liu Hongyu and, Zhang Xiaoliang - A Network Data Security Analysis Method Based on DPI Technology.

5) RehamTaher El-Maghraby, Nada MostafaAbdElazim and, Ayman M. Bahaa-Eldin - A Survey on Deep Packet Inspection.

6) T. Brinkhoff, "A framework for generating network-based moving objects, "GeoInformatica, vol. 6, no. 2, pp. 153-180, 2002.

7) G.-P. Roh, J.-W. Hwang and B.-K. Yi, "Supporting pattern-matching queries over trajectories on road networks, "IEEE Transactions on Knowledge and Data Engineering, vol, 23, no.11, pp. 1753-1758, 2011.

8) Cheng H, Yan X, Han J, Philip S (2008) Direct discriminative pattern mining. for effective classification.

9) Giannotti F, Nanni M, Pedreschi D - Trajectory pattern mining.

10) M Huang, P. Hu and L. Xia, "A grid based trajectory method for moving objects on fixed network", in Proceedings of the 18th International Conference on GeoInformatics, June 2010.

11) Coppi R, D'Urso P - Fuzzy K-means clustering models for triangular fuzzy time trajectories. Stat Methods Appl 11(1), pp. 22-40.

12) Gudmundsson J, Laube P, Wolle T (2012) Computational movement analysis. Springer handbook of geographic information, pp 725–741.