OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

SURVEY ON NETWORK TRAFFIC CLASSIFICATION TECHNIQUES WITH CORRELATION INFORMATION

Dheeraj Sudamrao Sadawarte [1], Shital Y Gaikwad[2]

ME Second year Student, Department of Computer Science and Engineering, Matoshri Pratishthans Group of Institutes, College of Engineering, Kupsurwadi, Nanded. (M.S.), India.[1]

Assistant Professor, Department of Computer Science and Engineering, Matoshri Pratishthans Group of Institutes, College of Engineering, Kupsurwadi, Nanded. (M.S.), India.[2]

sadawarte.dheeraj@gmail.com [1], shitalygaikwad@gmail.com [2]

Abstract: **Network Traffic classification is the need of today's emerging and rapidly growing computer network for network traffic analysis, network management, security monitoring, flow detection, QoS and lawful interception. It is possible to apply machine learning techniques to classify traffic based on flow statistical feature. Supervised and unsupervised classification algorithms have been applied to classify traffic. Conventional methods for classification include port based prediction and payload based deep inspection methods. In recent network environment the usual methods undergoes from some troubles like dynamic ports and encrypted applications. The nearest neighbor (NN) method having advantages, such as no need of training procedure, no risk of over fitting of parameters, and able to handle a large number of classes. But, the performance of NN classifier method affected if the size of training data is small. This paper conducts a survey on the various network traffic classification techniques, also focuses on a new non parametric traffic classification method, which makes use of correlation information for the purpose of classification.**

Keywords: *Traffic classification, network operations, security, nearest neighbor*

.----------∴∴∴.----------

I INTRODUCTION

There is a need to efficiently identify and classify network traffic into different classes plays a major role in network management, analyze behavior which will be useful in network security and monitoring [1][6]. Network classification is required to find out the network traffic characteristics and the overall traffic statistical flow behavior. Classifying network traffic shows considerable interest over last few years [6] [7].

Network traffic is the amount of data packets moving across a network at a given point of time. These packets are generated by one of the application or service running on source computer system, these packets is transmitted to application or service running on destination computer system. IP address is responsible for communication between two computers. Using port number a protocol or service or an application running on one computer system communicate with a protocol or a service or an application running on other system.

Classification of application generated traffic through traffic flows plays major role in network management and security. It is also used in network traffic analysis, network management, security monitoring, flow detection, QoS, intrusion detection and lawful interception [1].

Network traffic classification can be performed in two different ways. First is the packet level method, examines each packet's characteristics and application signatures. Second flow level methods are based on the aggregation of packets to flows and extraction of characteristics and statistical analysis from the flow.

Classification process needs features of a packet, like content from header or payload. Traffic can be classified by their generation application using well known port numbers, or by inspecting the contents of payload. The port-based prediction and payload based deep inspection are the conventional method for traffic classification [1]. Official assignments of port numbers for specific uses and their maintaining is the responsibility of Internet Assigned Number Authority (IANA). On the other hand, most of unofficial uses

both well-known and registered port numbers. The *well-known ports* or *system ports* are range from 0 to 1023 and the registered port numbers are ranges from 1024 to 49151. The IANA assigns port numbers for TCP or UDP in Port based attributes and also assigns source port and the destination port for every packet in the IP traffic.

All the applications in the network traffic do not have registered port numbers; hence it's very difficult to identify the unknown application using port based methods. Some applications dealing with dynamic port numbers so that it is difficult to classify such applications using port based techniques.

In Payload based deep inspection method, attributes are based on application layer level traffic signatures. Statistical based attributes related to traffic such as duration between the flow, packet ideal time, length of the packets and it's inter arrival time also play an important role in traffic classification. Payload based deep packet inspection technique match both the payload of the packet and known traffic signature but this method will not produce good classification accuracy when packets payload can be encrypted or obfuscated. So the conventional methods facing a number of problems, like the dynamic ports and contents of payload encrypted by applications.

Statistical based attributes relate to traffic statistical characteristics (e.g. flow duration, idle time, packets' inter-arrival time and length). These attributes are unique for certain classes of applications and enable distinguishing different source applications from one another.

Current research focuses on application of machine learning techniques applied on traffic classification based on flow statistical features. Machine learning can automatically search for and describe useful structural patterns in a supplied traffic data set, which is helpful to intelligently conduct traffic classification [6].

II OVERVIEW OF TRAFFIC CLASSIFICATION TECHNIQUES

Traffic classification using applications of machine learning techniques categorized as supervised methods or unsupervised methods.

A. Supervised Methods

In supervised traffic classification, there is general assumption that sufficient supervised training data is available. This supervised training data is analyzed by supervised traffic classification method and it produce an inferred function which can predict output class for any testing flow. Supervised learning induces knowledge structures that support the task of classifying new instances into pre-defined classes. Moore and Zuev [1] applied the supervised naive Bayes techniques to classify network traffic based on flow statistical features to address the problems suffered by payload-based traffic classification, such as encrypted applications and user data privacy. Williams, Zander, and Armitage [3] evaluate the supervised algorithms including naive Bayes with discretization, naive Bayes with kernel density estimation, C4.5 decision tree [2], Bayesian network, and naive Bayes tree. Nguyen and Armitage [6] proposed to conduct traffic classification based on the recent packets of a flow for real-time purpose. Auld, Moore, and Gull [5] extended the work of with the application of Bayesian neural networks [1] for accurate traffic classification.

B. Unsupervised Methods

The unsupervised methods (also called clustering) try to find cluster structure in unlabeled traffic data. Clustering can combine similar flow attributes using unsupervised learning. It assigns any testing flow to the application-based class of its nearest cluster. The use of clustering Algorithms for traffic classification is normally done in two phases. The first phase consists of training the model with a relatively small set of data (training data), and the second phase consists of using the trained model to classify unknown traffic. During the training phase, the training data is used to build clusters based on some criteria of similarity, which will ideally separate the data into similar clusters (groups). The second phase consists of assigning a class to the flows to be identified, depending upon the label of the cluster similar to each flow.

The Unsupervised Machine learning approach is based on a classifier built from clusters which are thus found and labeled in a training set of data. Once the classifier has been built, the classification process consist of the classifier calculating as to which cluster a connection is nearest to, and thereby using the label from the calculated cluster in order to recognize the connection. Zander [2] used AutoClass to group traffic flows and proposed a metric which is called intraclass homogeneity for cluster evaluation. Erman, Mahanti, and Arlitt [4], evaluated the k- means, DBSCAN and AutoClass algorithms for traffic clustering on two empirical data traces. The experimental research showed that traffic clustering can produce high-purity clusters when the number of clusters is set as much larger than the number of real applications. Generally, the clustering techniques [4] can be used to discover traffic from previously unknown applications. However, the clustering methods suffer from a problem of mapping from a large number of clusters to real applications.

III RELATED WORK

Nguyen and Armitage [6], presented a survey report on techniques for internet traffic classification using machine learning. New era of research community has started looking for IP traffic classification techniques that is independent of traditional techniques such as 'well known ports or registered ports' TCP or UDP port numbers, or reading the contents of packet header and payloads. Now the recent research work is

focusing on the use of statistical traffic flow characteristics to help in the identification and classification process, also focuses on emerging research into the application of Machine Learning (ML) techniques for classification IP traffic. Nguyen and Armitage [6] have reviewed 18 significant works, which provide context and motivation for the application of machine learning techniques to IP traffic classification.

As Zander, Nguyen, and Armitage, [2] reported their work 'Automated Traffic Classification and Application Identification using Machine Learning'**.** Classification and identification of network applications can be done using traffic flows which can be beneficial for network management and surveillance. Traditional classification methods depend on packet header fields or application layer protocol decoding. These methods have shortfalls like unpredictable port numbers and encrypted applications. To address above problem a novel method for traffic classification and application identification using an unsupervised machine learning technique is proposed.

Auld, Moore, and Gull [5], in "Bayesian Neural Networks for Internet Traffic Classification", paper explains, efficient internet traffic identification is an important tool for network management. It allows operators to better predict future traffic matrices and demands, security personnel to detect anomalous behavior, and researchers to develop more realistic traffic models. Here present a traffic classifier that can achieve a high accuracy across a range of application types without any source or destination host-address or port information. Auld, Moore, and Gull [5] suggested to use supervised machine learning based on a Bayesian trained neural network. However suggested technique uses training data with classes derived from packet content, training and testing were done using features derived from packet streams consisting of one or more packet headers. The suggested technique provides classification without access to the contents of packets; which offers wider application.

Jun Zhang, Chao Chen, Yang Xiang, Wanlei Zhou, and Yong Xiang [10] discuss a novel traffic classification approach to improve classification performance when very few training samples are available. Their work reported in "Internet Traffic Classification by Aggregating Correlated Naive Bayes Predictions". They proposed a new scheme; where discretized statistical feature describes traffic flows and flow correlation information is modeled by BoF (bag-of-flow). A new BoF-based traffic classification method is proposed to aggregate the naive Bayes (NB) predictions of the correlated flows. Also present an analysis on prediction error sensitivity of the aggregation strategies.

IV PROPOSED METHODOLOGY

A novel nonparametric approach is proposed to effectively incorporate flow correlation information into the classification process [11].

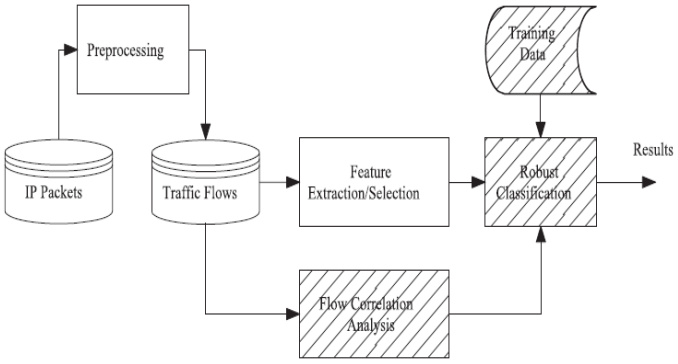

Figure 1 New Traffic Classification System Model (Source Jun Zhang et al. [11])

In the preprocessing, the system captures IP packets crossing a computer network and constructs traffic flows by IP header inspection or a data set can be applied to this phase. Flow is a group of sessions between two network addresses (IP pair) during the aggregation period. A new flow starts if the time between the end of a session (the last packet) and the start of a new session (first packet) is more than the defined idle time. The new session is then part of the new flow [9]. A flow consists of successive IP packets having the same five-tuple: (src ip; src port; dst ip; dst port; protocol) [8].

After that, each flow is represented by a set of statistical features that are extracted. Feature selection aims to select a subset of relevant features for building robust classification models. Flow correlation analysis is proposed to correlate information in the traffic flows. Finally, the robust traffic classification engine classifies traffic flows into application based classes by taking all information of statistical features and flow correlation into account. The novelty of our system model is to discover correlation information in the traffic flows and incorporate it into the classification process. Conventional supervised classification methods treat the traffic flows as the individual and independent instances.

Here "bag of flows" (BoF) used to model correlation information in traffic flows. A BoF consists of some correlated traffic flows which are generated by the same application.

A BoF can be described by

$$Q = \{x1,x2,...,xn\}$$

where xi is a feature vector representing the i^{th} flow in the BoF Q. The BoF Q explicitly denotes the correlation among n flows, {x1; . . . ; xn}.

Correlation Analysis

Here correlation analysis is conducted using a three-tuple heuristic [8], which can quickly discover BoFs in the real traffic data.

Three-tuple heuristic: in a certain period of time, the flows sharing the same three-tuple [9] {dst ip; dst port; protocol} form a BoF.

The correlated flows sharing the same three-tuple [9] [8] are generated by the same application. For example, several flows initiated by different hosts are all connecting to a same host at TCP port 80 in a short period. These flows are very likely generated by the same application such as a web browser.

Here different aggregation strategies can be applied, which have different meanings [11]

AVG-NN: combines multiple flow distances to make a decision for a BoF;

MIN-NN: chooses a minimum flow distance to make a decision for a BoF;

MVT-NN: combines multiple decisions on flows to make a final decision for a BoF.

V CONCLUSION

This paper surveys different approaches used for network traffic classification. The statistical feature based classification technique achieves better performance as compared to port based prediction method and payload based deep inspection method. The problem of traffic classification using very few supervised training samples is addressed a novel nonparametric approach, Traffic Classification using Correlation information (TCC), was proposed to investigate correlation information in real traffic data and incorporate it into traffic classification.

REFERENCES

[1] A.W. Moore and D. Zuev, "Internet Traffic Classification Using Bayesian Analysis Techniques," ACM SIGMETRICS Performance Evaluation Review (SIGMETRICS), vol. 33, pp. 50-60, June 2005.

[2] S. Zander, T. Nguyen, and G. Armitage, "Automated Traffic Classification and Application Identification Using Machine Learning," Proc. IEEE Ann. Conf. Local Computer Networks, pp. 250-257, 2005.

[3] N. Williams, S. Zander, and G. Armitage, "A Preliminary Performance Comparison of Five Machine Learning Algorithms for Practical IP Traffic Flow Classification," Proc ACM SIGCOMM, vol. 36, pp. 5-16, Oct. 2006.

[4] J. Erman, A. Mahanti, and M. Arlitt, "Internet Traffic Identification Using Machine Learning," Proc. IEEE Global Telecomm. Conf., pp. 1-6, 2006.

[5] T. Auld, A.W. Moore, and S.F. Gull, "Bayesian Neural Networks for Internet Traffic Classification," IEEE Trans. Neural Networks, vol. 18, no. 1, pp. 223-239, Jan. 2007.

[6] T.T. Nguyen and G. Armitage, "A Survey Of Techniques for Internet Traffic Classification Using Machine Learning," IEEE Comm. Surveys Tutorials, vol. 10, no. 4, pp. 56-76, Oct.-Dec. 2008.

[7] H. Kim, K. Claffy, M. Fomenkov, D. Barman, M. Faloutsos, and K. Lee, "Internet Traffic Classification Demystified: Myths, Caveats, and the Best Practices," Proc. ACM CoNEXT Conf., pp. 1-12, 2008.

[8] M. Canini, W. Li, M. Zadnik, and A.W. Moore, "Experience with High-Speed Automated Application-Identification for Network- Management," Proc. Fifth ACM/IEEE Symp. Architectures for Networking and Comm. Systems, pp. 209-218, 2009.

[9] Y. Wang, Y. Xiang, J. Zhang, and S.-Z. Yu, "A Novel Semi-Supervised Approach for Network Traffic Clustering," Proc. Int'l Conf. Network and System Security, Sept. 2011.

[10] Jun Zhang, Chao Chen, Yang Xiang, Wanlei Zhou, and Yong Xiang, "Internet Traffic Classification by Aggregating Correlated Naive Bayes Predictions," IEEE Transactions On Information Forensics and Security, vol. 8, no. 1, pp. 5-15, Jan 2013.

[11] Jun Zhang, Yang Xiang, Yu Wang, Wanlei Zhou, Yong Xiang, and Yong Guan, "Network Traffic Classification Using Correlation Information" IEEE Transactions On Parallel and Distributed Systems, vol. 24, no. 1, pp. 104-117, Jan 2013.