IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 2, Ver. III (Mar-Apr. 2014), PP 123-126
www.iosrjournals.org

Automated Packet Classification and Layer Identification of Network Packets a Review

Shancy P. Varghese, Jasmine Joseph
(Department of Computer Science and Engineering
Nehru College of Engineering and Research Center, Pampady, Thrissur,kerala)

Abstract: *Traffic classification is an automated process which categorizes computer network traffic according to various into a number of traffic classes. In this paper we are comparing different traffic classification methods and also presents an automated packet classification and layer identification to improve classification performance when few training data is available. In this we are grouping the similar flows and classifying it by using a classifier combination framework. We are aggregating naive Bayes (NB) predictions of the correlated flows. In this we are finding the denial of service (DOS) attack and also we are identifying which all applications are running in the systems which connected in that network. We are identifying the applications based on their port number.*

Keywords: *Traffic classification, network security, naive Bayes.*

I. Introduction

Traffic classification describes methods of classifying traffic based on features passively observed in the traffic, and according to specific classification goals. Traffic classification is an automated process which categorizes computer network traffic according to various parameters for example, based on port number or protocol into a number of traffic classes. Accurate network traffic classification is fundamental to numerous network activities, from security monitoring to accounting, and from Quality of Service to providing operators with useful forecasts for long-term provisioning. Real-time traffic classification has the potential to solve difficult network management problems for Internet service providers and their equipment vendors. Network operators need to know what is flowing over their networks promptly so they can react quickly in support of their various business goals.

Traditional traffic classification techniques may rely on the port numbers specified by different applications or the signature strings in the payload of IP packets. This cannot be supported due to dynamic port numbers and it may affect users' privacy policy. It is a big challenge for current network management is to handle a large number of emerging applications, where it is almost impossible to collect sufficient training samples in a limited time. In a complex network situation it is difficult to obtain a high performance using small set of training data. Traffic labeling is also time consuming for new and encrypted applications. The objective of this paper is it presents a novel traffic classification scheme to improve classification performance when few training data are available. In the proposed scheme, traffic flows are described using the discretized statistical features and flow correlation information is modeled by bag-of-flow [1].

The remainder of the paper is structured as follows. Section II reviews some related works. The traffic classification scheme is proposed in Section III. Finally, the paper is concluded in Section V.

II. Related Work

Classifying traffic flows according to the applications that generate them is an important task for effective network planning and design, and monitoring the trends of the applications in operational networks. However, an accurate method that can reliably identify the generating application of a flow is still to be developed. Blinc is based on observing and identifying patterns of host behavior at the transport layer. Approach has two important features. First, it operates in the dark, having no access to user payload is possible, well-known port numbers cannot be assumed to indicate the application reliably, and only use the information that current flow collectors

provide. These restrictions respect privacy, technological and practical constraints. Second, it can be tuned to balance the accuracy of the classification versus the number of successfully classified traffic flows [2]. Disadvantage of BLINC is it cannot identify specific application sub-types this technique is capable of identifying the type of an application but may not be able to identify distinct applications. Another disadvantage is that of encrypted transport layer headers layer-3 packet headers be also encrypted, methodology cannot function. Also network address translators have to handle most cases.

Another approach is to apply a Naive Bayes estimator to categorize traffic by application. This approach capitalizes on hand-classified network data, using it as input to a supervised Naive Bayes estimator. In this it illustrate the high level of accuracy achievable with the Naive Bayes estimator. Accurate network traffic classification is fundamental to numerous network activities, from security monitoring to accounting, and from Quality of Service to providing operators with useful forecasts for long-term provisioning. Supervised Machine-Learning is used to classify network traffic. Supervised learning is machine learning task of inferring a function from a labeled training data. Uniquely, data that has been hand-classified based upon flow content to one of a number of categories. Sets of data consisting of the hand-assigned category combined with descriptions of the classified flows are used to train the classifier [3]. The disadvantages of this work are there is no best possible discriminator and lack of feature discretization.

The task of identifying the optimal set of flow attributes that minimizes the processing cost, while maximizing the classification accuracy. The dynamic classification and identification of network applications responsible for network traffic flows offers substantial benefits to a number of key areas in IP network engineering, management and surveillance. This proposes a novel method for traffic classification and application identification using an unsupervised machine learning technique. Flows are automatically classified based on statistical flow characteristics. Evaluate the efficiency of approach using data from several traffic traces collected at different locations of the Internet. Use feature selection to find an optimal feature set and determines the influence of different features [4]. The main disadvantage is that the accuracy is only 86%. Another one is quantify the performance in terms of processing time and memory consumption and to investigate the trade-off between the approach's accuracy and processing overhead.

Laurent et al. [5] propose a technique that relies on the observation of the first five packets of a TCP connection to identify the application. The early detection of the applications associated with TCP flows is an essential step for network security and traffic engineering. Enterprise or campus networks usually impose a set of rules for users to access the network in order to protect network resources and enforce institutional policies for instance, no sharing of music files or gaming. This leaves network administrators with the daunting task of identifying the application associated with a traffic flow on the fly and controlling user's traffic when needed. Therefore, accurate classification is essential. The main disadvantage of this method is that if the packet is received out of order it cannot be processed.

A flow classification mechanism based on three simple properties of the captured IP packets: their size, inter-arrival time and arrival order, Traffic classification mechanisms belongs to a wide set of tools That helps the allocation, control and management of resources in TCP/IP networks, and improve the reliability of Network Intrusion Detection Systems. An effective mechanism for the classification of the traffic flows according to the application layer protocols that generated them can suggest suitable measures to prevent or ease network congestion, to deploy QoS–aware mechanisms successfully, or to counter network attacks. This approach belongs to yet another class of techniques, those which try to classify network traffic relying exclusively on the statistical properties of the flows [6]. The main disadvantage of this approach is that data has to be trained and also the accuracy. The technique can determine with a relatively small error ratio the application protocol behind network flows, at least with a reduced set of protocols, when the classifier has been properly trained.

The table 2.1 shows the comparison of traffic classification methods. It includes adaptability, Real-time classification, weather it is able to classify the flows in progress and detectability. Adaptability refers to weather the approach is adaptive to changing or different traffic characteristics. Detectability means detection ability of new or anomalous applications. Real time classification indicates weather it is possible to classify in real time and to classify the traffic as it comes is denoted by classify the flows in progress.

Method	Adaptability	Real-time Classification	Classify flows in progress	Detectability
Blinc	Retuning needed	No	Not Addressed	Yes
Bayesian Analysis Technique	May need to retrain	No	Yes	No
Automated Traffic Classification	May need to retrain	NA	NA	Yes
Traffic Classification On The Fly	No	No	Yes	No
Statistical Fingerprinting	No	Yes	Yes	No

Table 2.1: Comparison of traffic classification methods

III. Classification Scheme

This section presents a novel NB-based classification scheme to deal with the correlated flows in an effective way, which can significantly improve the classification performance even with a small set of supervised training data. Also detecting the Dos packets and identifying the application running in the systems in the network.

A. Naive bayes predictions: A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes therom with strong independence assumptions. A more descriptive term for the underlying probability model would be independent feature model. Naive Bayes belongs to a group of statistical techniques that are called supervised classification as opposed to unsupervised classification. In simple terms, a naive Bayes classifier assumes that the presence or absence of a particular feature of a class is unrelated to the presence or absence of any other feature, given the class variable. Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood, in other words, one can work with the naive Bayes model without believing in Bayesian probability or using any Bayesian methods.

B. Classification process: The classification process focus on flow-level traffic classification. In the preprocessing, the system captures IP packets crossing a target network and constructs traffic flows by checking the headers of IP packets. A flow consists of successive IP packets with the same 5-tuple: source IP, source port, destination IP, destination port, and transport layer protocol. Apply a heuristic way to determine the correlated flows and model them using bag-of-flows. If the flows observed in a certain period of time share the same destination IP, destination port, and transport layer protocol, they are determined as correlated flows and form a BoF. For the classification purpose, a set of flow statistical features are extracted and discretized to represent traffic flows. A novel approach is proposed for traffic classification, namely aggregation of correlated NB predictions, which consists of two steps. In the first step, the single NB predictor produces the posteriori class-conditional probabilities for each flow. In the second step, the aggregated predictor aggregates the flow predictions posteriori probabilities to determine the final class for BoF.

C. A bof-based classification framework: In the proposed scheme, a set of correlated flows are generated by the same application, which is modeled using a bag of flows, $X=\{x_1,\ldots..,x_b\}$. Since the flows,$x_1,\ldots..,x_b$, belong to the same application-based class, such correlation information can be utilized to improve the classification results. Therefore, aim to aggregate the individual predictions of the correlated flows so as to conduct more accurate classification. The research shows that the goal can be achieved by following the approach of classifier combination. The BoF-based classification can be fitted into Kittler's theoretical framework for classifier combination. The classifiers are combined to form the aggregated predictor, which improves the classification accuracy.

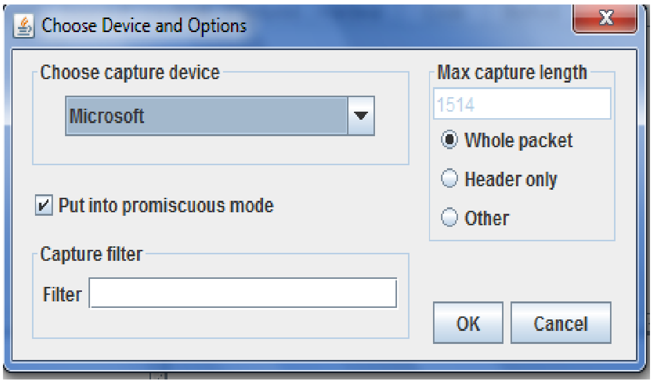
D. Aggregation of correlated nb predictions: The simple predictor is unstable due to a small set of training data. The main aim of this method is to classify the traffic with fewer amounts of data. So, the aggregation of correlated flow predictions can improve the performance to generate the aggregated predictor. BoF-based traffic classification is solved by aggregating correlated NB predictions. That is we get more accurate results while doing the internet traffic classification. In this we are using single NB predictor and aggregated NB predictor

Single NB Predictor**:** NB algorithm to produce a set of posterior probabilities as predictions for each testing flow. It is different to the conventional NB classifier which directly assigns a testing flow to a class with the maximum posterior probability. Considering correlated flows, the predictions of multiple flows will be aggregated to make a final prediction. Naive Bayes classifier is chosen for this scheme due to two reasons. Firstly, it has demonstrated high classification speed and good performance using the discretized statistical features in traffic classification. Secondly, it is easy for naive Bayes classifier to produce the posterior probability that a testing flow belongs to a traffic class.

Aggregated Predictor: Under Kittler's theoretical framework, a number of combination methods can be derived from the Bayesian decision theory which can be used for aggregated predictor. Use sum rule to aggregate the classifiers. In this approach, BoF-based NB, to aggregate correlated NB predictions in this work, which results in a more accurate aggregated predictor for traffic classification.

E. Dos detuction and identifying applications: Here we are detecting the dos packets. The server will respond all the packets which are coming to it. So the hacker or intruder who wants to decrease the efficiency of the server will send dummy packets, which contains no information. So to avoid that we are setting a threshold value and if the value of the packet is greater than the threshold we are marking it has attack packet. We can also identify the applications which are running in the systems which are connected to this network. This is based on the port number. According to the port number we are identifying which application is running.

The Fig 3.2 shows the choose devise and options window. This window will appear first when we run the program. We are selecting the type of connection through this window. Also we can specify whether we need full packet or only the header or the size of data we want to retrieve. We have to mark the promiscuous mode because normally a network interface will only receive packets directly addressed to the interface. Promiscuous mode allows the interface to receive all packets that it sees weather they are addressed to the interface or not.



Choose capture device: Microsoft

☑ Put into promiscuous mode

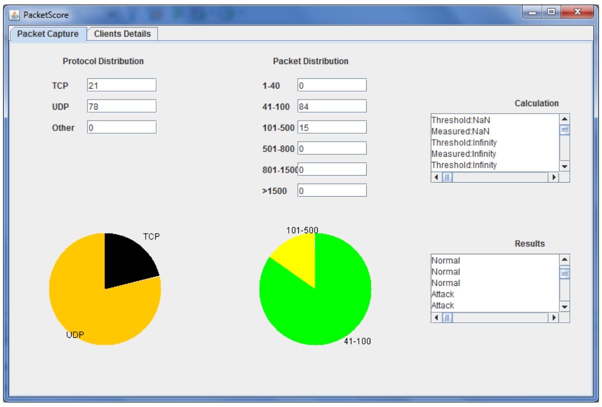
Capture filter
Filter:

Max capture length: 1514

● Whole packet
○ Header only
○ Other

Fig3.1 Choose Devise and Options

The Fig3.1 shows the classification window. In the figure the protocol distribution describes the classification of received packets. In this the packets are classified as tcp packet, udp packets and others. The result of classification is also shown in the form of a pie graph. The packet distribution shows that the classification of packets according to their size. That is we are grouping the packets to different groups according to their sizes. result of classification is also shown in the form of a pie graph. In the calculation window it showing the value obtained after calculating. To detect weather the incoming packet is a normal packet or an attack packet. The result window shows the type of incoming packet. That is weather the incoming packet is attack or not.



Protocol Distribution

TCP 21
UDP 78
Other 0

Packet Distribution

1-40 0
41-100 84
101-500 15
501-800 0
801-1500 0
>1500 0

Calculation

```
Threshold:NaN
Measured:NaN
Threshold:Infinity
Measured:Infinity
Threshold:Infinity
```

Results

```
Normal
Normal
Normal
Attack
Attack
```

Fig4.1Classification window

IV. Conclusion

This traffic classification scheme can effectively improve the classification performance and also it can detect the dos packets and identify the applications that generate them. We are finding the Dos packets based on the threshold value calculated and identifying the applications based on the port numbers that generate them. This method gives solution to high performance traffic classification without time consuming.

References

[1] Jun Zhang, Chao Chen, Yang Xiang, Wanlei Zhou, and Yong Xiang, “Internet traffic classification by aggregating naïve bayes prediction,” in Proc. SIGCOMM Comput. Commun. Rev., Jan. 2013, vol. 8, pp. 5–18.

[2] T. Karagiannis, K. Papagiannaki, and M. Faloutsos, “BLINC: Multi- level traffic classification in the dark,” in Proc. SIGCOMM Comput. Commun. Rev., Aug. 2005, vol. 35, pp. 229–240.

[3] A. W. Moore and D. Zuev, “Internet traffic classification using bayesian analysis techniques,” in SIGMETRICS Perform. Eval. Rev., Jun. 2005, vol. 33, pp. 50–60.

[4] S. Zander, T. Nguyen, and G. Armitage, “Automated traffic classifi- cation and application identification using machine learning,” in Proc. Ann. IEEE Conf. Local Computer Networks, Los Alamitos, CA, 2005, pp. 250–257.

[5] L. Bernaille, R. Teixeira, I. Akodkenou, A. Soule, and K. Salama- tian, “Traffic classification on the fly,” in Proc. SIGCOMM Comput. Commun. Rev., Apr. 2006, vol. 36, pp. 23–26.

[6] M. Crotti, M. Dusi, F. Gringoli, and L. Salgarelli, “Traffic classification through simple statistical fingerprinting,” in Proc. SIGCOMM Comput. Commun. Rev., Jan. 2007, vol. 37, pp. 5–16.

[7] T. T. Nguyen and G. Armitage, “A survey of techniques for internet traffic classification using machine learning,” Commun. Surveys Tuts., vol. 10, no. 4, pp. 56–76, 4th Quarter 2008.

[8] Y.-S. Lim, H.-C. Kim, J. Jeong, C.-K. Kim, T. T. Kwon, and Y. Choi, “Internet traffic classification demystified: On the sources of the dis- criminative power,” in Proc. 6th Int. Conf., Ser. Co-NEXT’10, New York, 2010, pp. 9:1–9:12, ACM