TRAFFIC CLASSIFICATION FOR UNKNOWN FLOWS

[1]V.R. Meenachi, [2]K.M. Kirthika, [3]S.J. Savitha, [4]D. Betteena Sheryl
[1, 2, 3, 4]Research Scholars, Sri Ramakrishna Institute of Technology
meenachi.cse@srit.org

Abstract: Traffic classification encounters more critical problems in current networks. The network does not broach the subject of unknown traffic classification mainly due to the number of unknown applications and lack of supervised information, thus limiting the benefits of statistical features pertaining to a network flow and machine learning techniques. In this paper to overcome the problem of traffic classification the techniques such as flow label propagation and compound classification is utilized. Flow label propagation which can automatically and accurately label the unlabeled flows enhances the capability of the network by using Nearest Cluster based Classifier (NCC) technique. The compound classification method combines a number of flow predictions to make more accurate classification of the network flows. The above techniques when applied provided better results when compared with other existing techniques.

Keywords: flow label propagation, compound classification.

1. INTRODUCTION

Network traffic is defined as the number of packets sent/received over the network. It monitors the packets. Traffic Measurement is the process of measuring the features of traffic on a network the process of measuring the amount and type of traffic on a particular network is known as traffic measurement. Traffic classification plays a major role in modern network security and architectures. For instance, traffic classification is normally an essential component in the products used for QoS control and intrusion detection. With the cloud computing techniques increase in the amount of application is more with encryption techniques. This situation makes it difficult to classify traffic flows based on their generation applications. Traditional traffic classification techniques depend on the specific port numbers used by application, or by the payload of IP packets. The unsupervised technique confront a number of problems in the recent network such as dynamic port numbers, data encryption and user privacy protection.

1.1 Network traffic classification

Traffic classification is a process which categorizes traffic of computer network based on various parameters, for example, based on the port number or protocol, the number of traffic classes is classified.
Many supervised and unsupervised algorithm has been applied in traffic classification. In supervised traffic classification, the labeled training samples of predefined traffic class are learned with the flow classification model. The supervised methods classify any flows into predefined traffic classes, so that they cannot deal with unknown flows generated by unknown applications. For high accuracy sufficient labeled training data of supervised method is needed. By variation, the clustering-based methods can automatically group a set of unlabeled training samples and apply the clustering results to construct a traffic classifier. To attain high purity number of clusters has to be set large. It is difficult to map from large to small cluster in real time application without supervised method.

1.2 Classification methods

- **Port numbers:** In this method, the port numbers used for network communication is identified and the load is calculated.
- **Data Packet Inspection:** It is a form of computer network packet filtering which examines the packet header, searching for protocol which is of non-compliance, viruses, intrusions, spam or defined criteria to decide whether the packet may pass or if it needs to be routed to different destination or for the purpose of collecting statistical information.

Statistical classification: It includes machine learning techniques such as supervised learning, unsupervised Learning, semi Supervised learning which works in

categorizing the data based on training set of data containing known and unknown instances.

2. RELATED WORK

2.1 Security Issues in Network with Internet Access

Armbrust et. al. [1] describes the basic principles of designing and administering a relatively secure network. The principles are illustrated by describing the security issues in a hypothetical company faces as the networks that support its operation. To a final state in which the Internet is finally integrated into its operations and the company participates in international electronic commerce. At each stage, the vulnerabilities and threats of the company is noted as well as the counter measure for that threat and risk factor is also analyzed. Network security policy, services and Internet architecture and vulnerabilities provides additional technical detail underlying the scenario is discussed. Lastly, a number of building blocks for secure networks are presented that can mitigate some of the vulnerabilities.

2.2 Change-Point monitoring for the detection of Denial of Service (DoS) Attacks

Roughan et. al. [2] presents a simple and robust mechanism to detect denial of Service(DoS), called Change-Point Monitoring (CPM). The key of CPM is based on the inherent behaviors of network protocol and is an instance of the Sequential Change Point Detection. To make the detection mechanism and traffic patterns hard, a nonparametric Cumulative Sum (CUSUM) method is applied, thus making the detection mechanism robust, more generally applicable and easier for deployment. CPM does not require all information, only a few variables to record the protocol behaviors. The statelessness and low computation overhead of CPM make it immune to any flooding attacks. The evaluation results show that CPM has short detection latency and high detection accuracy to find flows of the network.

2.3 Scalable Attack Detection in the Network

At present intrusion detection and prevention systems pursue to detect a various class of network intrusions (e.g., DoS attacks, worms, port scans) at network standpoints. Woefully, many IDS systems keep per-connection or per-flow state to detect malicious TCP flows. Thus, it is surprising that the IDS systems do not scaled to multi-gigabit speeds. By contrast, both router lookups and fair queuing have scaled to high speeds using aggregation. Kim et.al [3] initiated research into the question as to whether one can detect attacks without keeping per-flow state. Such aggregation, while making fast implementations possible, causes two problems. First, aggregation can cause behavioral aliasing where, for example, good behaviors can cause to look like bad. Second, aggregated schemes are capable of spoofing by which the intruder sends attacks that have aggregate behavior. In addition to existing approaches for scalable attack detection, this work proposed a narrative data structure called partial completion filters (PCFs) that can detect claim-and-hold attacks in the network.

2.4 Advanced and authenticated marking schemes for IP Trace back

Securing against distributed denial-of-service attacks is the hardest security problems on the Internet today. One difficulty to thwart these attacks is to trace the attack source because they often use invalid or spoofed IP source addresses to disguise the true origin. Nguyen and Armitage [4] presented two new schemes, the advanced marking scheme and the authenticated marking scheme, which allow the dupe to trace-back the origin of spoofed IP packets. These techniques feature low network and router overhead, and support incremental deployment.

3. PROPOSED METHODOLOGY

A system model as shown in fig 1 to incorporate flow correlation into a semi-supervised method, which possesses the capability of unknown flow detection

The flow label propagation is used to automatically label relevant flows from a large unlabeled dataset in order to address the problem of small supervised training set. The compound classification is used to jointly identify the correlated flows in order to further boost that classification accuracy.

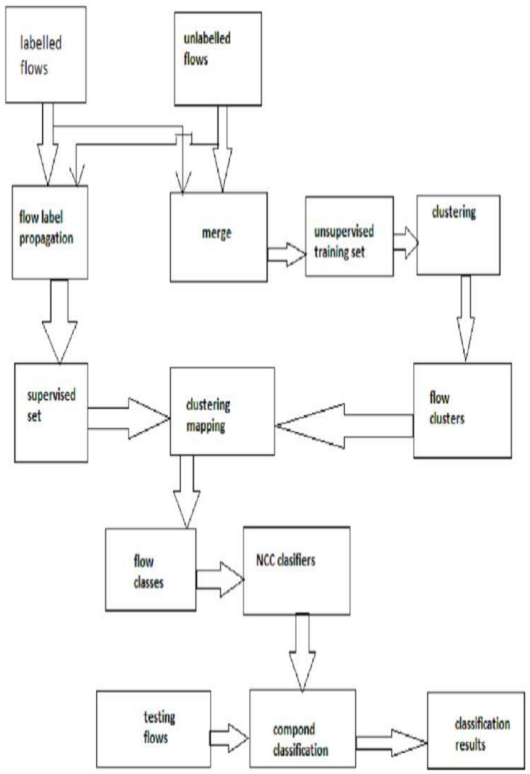

Figure 1: System model

3.1 SYSTEM MODEL

A novel system model represented in fig 1is with the capability of detecting unknown flows generated by unknown applications, the proposed system model can utilize flow correlation to effectively improve traffic classification. At the unsupervised training stage, a small number of labelled traffic flows and a large number of unlabeled traffic flows are combined to constitute an unsupervised training data set for traffic clustering. Meanwhile, flow label propagation extends the labelled traffic set by automatically searching for flows which are correlated to the pre-labelled flows in the unlabeled traffic set. Then, the clusters, as the output of traffic clustering on the unsupervised training data, are mapped into the application-oriented traffic classes with the assistance of labelled flows. The categorized traffic flows are used to train a traffic classifier such as nearest neighbor. In the unsupervised testing stage, compound classification on the correlated flows instead of classifying individual traffic flows are performed.As a semi-supervised learning, system model involves a small number of pre-labelled traffic flows and a large number of unlabeled traffic flows in the training stage. Unknown applications can be handled by assigning

their flows to unknown clusters. Since the number of pre-labelled traffic flows is small, it will cause two problems, false unknown clusters and weak traffic classifier. Flow label propagation is proposed to extend the labelled traffic set, so as to significantly reduce the amount of clusters which are related to known applications but inaccurately labelled as unknown in the training stage. Compound classification is proposed to classify the correlated flows instead of classifying individual traffic flows in the testing stage, which can further boost the classification accuracy.

4. EXPERIMENTAL RESULTS

The traffic classification scheme is implemented in a Local Area Network of 30 nodes. The dataset is collected using Wire shark network stimulator.

Table 1: Sample Dataset

Protocol	Source port	New column	Label
TCP	13703	1	FTP
TCP	13703	2	DNS
UDP	50312	3	???
ARP	13703	4	???
UDP	13703	5	HTTP
UDP	62446	6	SMTP

The known are classified by the protocol names and unknown by ??? in label column as shown in table 1. The collected dataset has to be browsed and uploaded in the database. Two types of datasets are used, one for labeled flow and another one for unlabeled flow. In flow label propagation, the labeled flow and unlabeled flow tables are merged and a new table called flow label is generated. The k-means algorithm is used to

cluster the data. After applying k-means, the dataset will be represented as shown in table 2.

Table 2: Cluster classification

Protocol	Source port	New column	label	Cluster
TCP	13703	1	FTP	c1
TCP	13703	2	DNS	c2
UDP	50312	3	???	c3
ARP	13703	4	???	c4
UDP	13703	5	HTTP	c5

In compound classification module, a set of tuples are taken and are applied with the composition rule to cluster into 10 tables. Both nearest neighbor classification and compound classification are combined to identify the unknown flow. After this unknown flow will be stored in one table, the protocol is classified by entering the column number and source port number from new column of unknown flow which is present in the dataset. The protocols of same cluster are grouped separately as shown in table 3.

Table 3: Compound classification

c1	c2	c3	c4
FTP	SMTP	DNS	HTTP
FTP	SMTP	DNS	HTTP
FTP	SMTP	DNS	HTTP

The flow label propagation is one reason why the proposed approach works so well when only very few supervised samples are available. The other reason is of compound classification which can jointly classify correlated flows more accurately. In fact, conventional supervised methods perform very badly even without unknown traffic if the size of supervised training set is too small. Since the flow label propagation is independent to classification algorithms, in the future, a pre-processing step with any supervised methods can be used in order to increase the size of the supervised training set. However, it should be pointed out that the key concern of this project is unknown traffic. The flow label propagation cannot deal with unknown traffic straightforward. A semi supervised scheme by combining flow label propagation and compound classification to effectively handle unknown traffic.

5. PERFORMANCE ANALYSIS

Fig.2, indicates the performance of classification for unknown flows

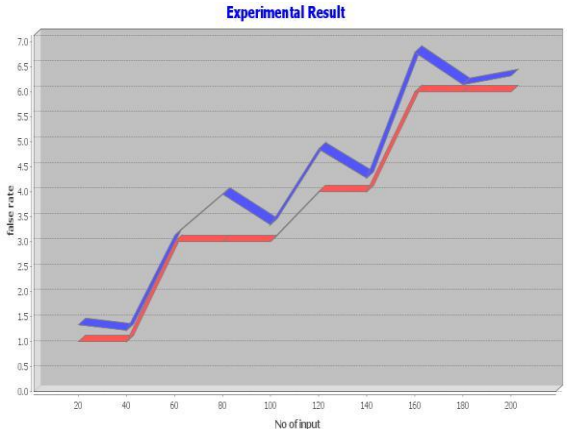

Figure 2: Classification Performance

Instead of testing each data into groups which will not accurately label the data, testing individual data and classifying into tuples to increase the performance. Tuples of same protocol are identified separately and with the help of clustering unknown data are labeled. A large number of experiments are done to comprehensively evaluate the proposed method. The proposed method is compared to five state-of-the-art methods for traffic classification: C4.5, k-NN, Bayes Network, Naive Bayes (NB), and Erman's semi-supervised method as shown in fig 2 simulate the problem of unknown applications; manually set some identified applications are set as unknown in the experiments.

6. CONCLUSION

Traffic classification confronts more critical problems in current advanced network and system, especially in cloud computing environment. In this paper, a novel traffic classification method to address the problem of unknown applications in the crucial situation of small supervised training data. The proposed method introduces two new techniques to sufficiently utilize flow correlation information. One is flow label propagation which can automatically accurately label more unlabeled flows to enhance the capability of nearest cluster based classifier (NCC). The other is compound classification which can combine a number of flow predictions to make more accurate classification. A large number of experiments were carried out on traffic datasets to evaluate the proposed method. The results showed that the proposed method outperforms five state-of-the- art traffic classification methods including C4.5, kNN, Naive Bayes, Bayesian Network, and Erman's semi supervised method. During more specific comparison with Erman's semi supervised method, the proposed method displayed more robust ability to various parameters and superior unknown detection performance especially on false detection.

REFERENCES

[1] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "A view of cloud computing," Communication, ACM, vol. 53, pp. 50–58, Apr. 2009.

[2] M. Roughan, S. Sen, O. Spatscheck, and N. Duffield, "Class-of-service mapping for QoS: A statistical signature-based approach to IP traffic classification," in Processing, ACM SIGCOMM Conference on InternetMeasurement, pp.135–148.2010.

[3] H. Kim, K. Claffy, M. Fomenkov, D. Barman, M. Faloutsos, and K. Lee, "Internet traffic classification demystified: myths, caveats, and the best practices," in Processing, ACM CoNEXT Conference, pp. 1–12,2009

[4] T. T. Nguyen and G. Armitage, "A survey of techniques for internet traffic classification using machine learning," IEEE Communication Surveys& Tutorials, vol. 10, no. 4, pp. 56–76, Fourth Quarter 2011.

[5] Jun Zhang, Chao Chen, "An Effective Network Traffic Classification with Unknown Flow Detection," IEEE Transactions on Network and Service Management, vol. 10, no. 2, June 2013.

[6] W. Moore and D. Zuev, "Internet traffic classification using Bayesian analysis techniques," SIGMETRICS Performance Evaluation, vol. 33, pp. 50– 60, June 2010.

[7] M. Canini, W. Li, M. Zadnik, and A. W. Moore, "Experience with highspeed automated application-identification for network-management," inProcessing, 2009 ACM/IEEE Symposium on Architectures for Networking and Communications Systems, pp. 209–218.

[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 discriminative power," in Processing, 2010 International Conference, pp. 9:1–9:12.

[9] S. Valenti, D. Rossi, M. Meo, M. Mellia, and P. Bermolen, "Accurate,fine-grained classification of P2P-TV applications by simply countingpackets," in Processing, 2009 International Workshop on Traffic Monitoring and Analysis, pp. 84–92.

[10] A. Finamore, M. Mellia, and M. Meo, "Mining unclassified traffic using automatic clustering techniques," in Processing, 2011 TMA International Workshop on Traffic Monitoring and Analysis, pp. 150–163