3C4-IOS-4a-4

# Internet Traffic Classification using Multi-Classifier Systems

Amjad Twalo [*1] Tsuyoshi Murata

Tokyo Institute of Technology

The field of Internet traffic classification has been growing fast recently. This growth is based on the increasing number of Internet users, and varieties of data intensive applications being used, such as video streaming and file sharing services. We have used an already implemented classification platform (Traffic Identification Engine - TIE) to implement ensemble classifiers and tried different configurations of classifiers, traffic features, and decision combiners in order to enhance classification accuracy. For this purpose, we gathered and analyzed anonymous traffic data with its corresponding ground truth to be used for the training and testing of the proposed methods.

## 1. Introduction

Traffic classification is a method for assigning classes to network flows based on features observed in the traffic. This field of research has gained a lot of attention recently for a number of reasons. Some of them are: (1) The explosion in the volume of data going through network nodes, and the proliferation of high speed internet access and the wide spread of data intensive applications, which created a need for Internet Service Providers to fine tune their network configurations in order to accommodate this amount of different traffic types and provide better quality of service for their customers; (2) The file sharing and copy right laws controversy between the file sharing community and the intellectual properties representatives [1]; (3) The increase in the number and sophistication of network attacks, which motivated the search for better and smarter methods for mitigating these threats.

In our paper, we propose a traffic classification method that uses multiple classifiers to achieve classification accuracy higher than each single classifier. We used a "*mixture of experts"* multi-classifier system in which a separate classifier is trained to determine the participation of each classifier in the final classification decision. We will describe our proposed system and show our preliminary results.

The remainder of this paper is organized as follows. After reviewing the related work in section 2, we will proceed to describe our method in section 3. In section 4, we will describe our data collection and the classification platform that was used. We discuss our results in section 5, and section 6 concludes the paper.

## 2. Related work

There are three main approaches to classify Internet traffic.

**Payload-based** approach**:** In this method, the content of the packet's payload (the section of the packet which contains the actual data generated by the user) is read and compared to a database of payload signatures in order to identify the application that generated the packet. This method produces highly accurate classification, but it is very computationally expensive, and requires access to the payload, which is not permitted in most of practical scenarios, as it would be an invasion of privacy. Nonetheless, this method is a very good tool to generate ground truth data which is used to train other classification algorithms.

**Port-Based** approach**:** This method assumes that each port is associated with one protocol or application. The classification is done by reading the port number in the packet header, and looking it up in a list of commonly used port numbers (such as IANA list of assigned port numbers [10]). The port-based method has been proven to be unreliable [6], due to a substantial number of applications using ports different from their registered ports, or using random ports not registered in the IANA registry, and in some cases, IP layer encryption is applied to the TCP header, rendering it impossible to read the port numbers.

**Statistical features** approach: This method uses machine learning algorithms such as SVMs, neural networks, decision trees or any different machine learning techniques to classify network traffic using as input, statistical features of the traffic flows, such as the number of packets, packet sizes, inter arrival time and flow duration. Extensive literature has been produced applying machine learning and data mining techniques to classify Internet traffic data. Nguyen *et al.* [2] categorizes and reviews these studies in term of the choice of machine learning algorithms implemented, and contribution to the field.

A recent trend in the literature is a multi-classifier approach, in which the premise is that, by combining the results of different base classifiers, we might obtain better classification accuracy. The assumption is that the consensus of a set of classifiers would compensate for the shortcomings of one single classifier [7]. In [5] Dainotti *et al.* implemented a multi-classifier system in which several classification algorithm were used in parallel, and their results were combined using different combination algorithms such as majority voting, Naïve Bayes and Dempster-Schafer combiner. Their paper concludes that this approach can possibly enhance both classification accuracy and early classification of the traffic flows.

The methods mentioned above rely on combining the classification decision of different classifiers to come to the final decision i.e. classifier fusion. In this paper, we instead propose another multi-classifier approach, using classifier selection instead of classifier fusion.

[*1]Amjad Twalo, Tokyo Institute of Technology, W8-59 2-12-1 Ookayama, Meguro, Tokyo 152-8552, twalo.a@ai.cs.titech.ac.jp

## 3. Method

There are two main strategies in combining classifiers: fusion and selection [3]. In classifier fusion, each ensemble member is supposed to have knowledge of the whole feature space. Whereas in classifier selection, each ensemble member is supposed to know well a part of the feature space and be responsible for objects in this part. This can particularly useful in traffic classification where not all classifiers work well with all of the features. That means that we have to identify the subset of features that can identify which classifier performs better on the given data.

In our study, we used an ensemble method called *mixture of experts.* In this method we make use of a separate classifier, which determines the participation of the experts (classifiers) in the final classification decision. Figure 1 illustrates basic structure of the model we used.

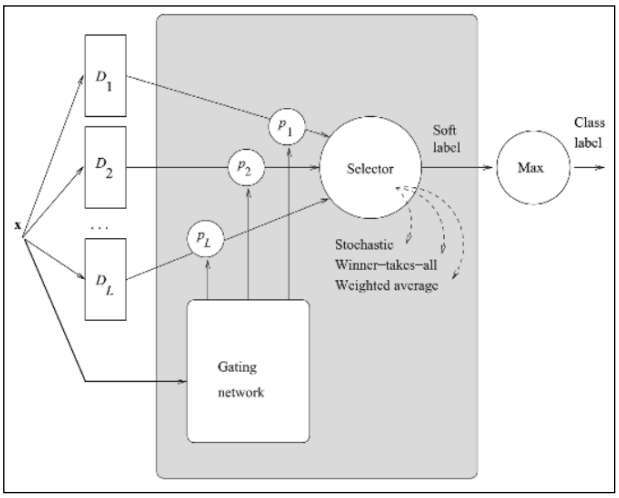


**Fig. 1** Basic Structure of Mixture of Experts System [3]

$\boldsymbol{X}$ is the flow features.

$\boldsymbol{D_1,D_2…D_L}$ are the base classifiers.

$\boldsymbol{P_1,P_2…P_L}$ are the weights generated by the gaiting network.

The gaiting network that we used is a simple perceptron that takes as input the sizes of the first 20 packets of the flow, and then produces output as the weights $P_1…$ $P_L$. The selector chooses the classifier using the weights and a selection strategy. In our experiment we used "winner-takes-all", where the selector chooses the classifier with the highest weight and discards the rest.

## 4. Data set and classification platform

### 4.1 Data set collection

We collected our dataset in a home Ethernet network during the span of 4 days. The traffic trace's size was 12.4 GB. The trace consisted of 17 million packets and over 180k biflows (bi-directional traffic flows between two ip addresses generated by the same application). Figure 2 shows the data collection setup.

We used TCPDUMP [9] utility to capture traffic at the Ethernet port and dump it into 10 minutes long traces. These traces would be classified using a deep packet inspection (DPI) classifier using the TIE platform [4] with only the L7 plugin [8] enabled. After that, the trace is anonymized using TCPANON [11].

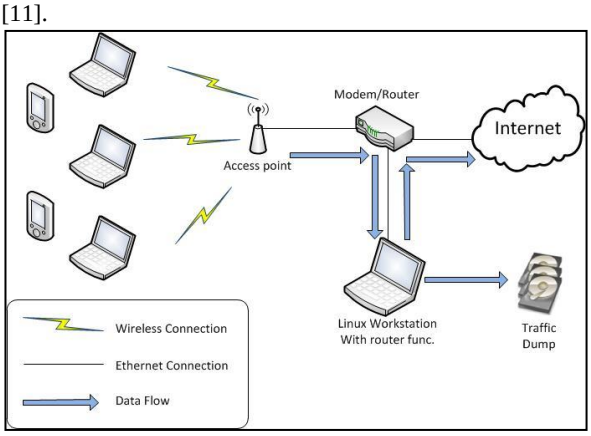


**Fig. 2** Data Collection Set up

This anonymized trace is then used as training data for the multi-classification system, and the DPI classification results are used as the corresponding ground truth data. Figure 3 illustrates the data set compiling set up.

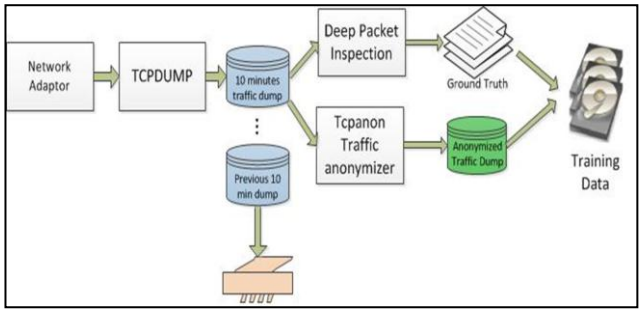


**Fig. 3** Data set compiling set up

After compiling the training data and ground truth for each 10 minutes trace, we delete the original trace thus leaving no personal data in storage. Table 1 shows the traffic breakdown in the data set.

**Table 1** Data set traffic breakdown

| Application | % of Biflows |
|---|---|
| Bittorrent | 43.12 |
| Http | 26.68 |
| EDONKEY | 9.6 |
| Skype | 6.92 |
| DNS | 3.35 |
| SMTP | 1.17 |
| SSL | 0.56 |
| Other | 8.6 |

### 4.2 Classification platform

We use in our study a classification tool called Traffic Identification Engine (TIE) [4]. This platform allows for fair evaluation and comparison of different techniques. It is a modular multi-classifier system that allows for flexibility in implementing classifiers and decision combination strategies.

We exported classification features from TIE into ARFF files compatible with WEKA platform [12], which is a suite of already implemented general purpose machine learning algorithms. Classification results from each classifier are then returned to TIE and combined using our newly implemented decision combiner using the "*mixture of experts"* classifier selection method. Figure 4 outlines the structure of TIE, combined with WEKA platform.

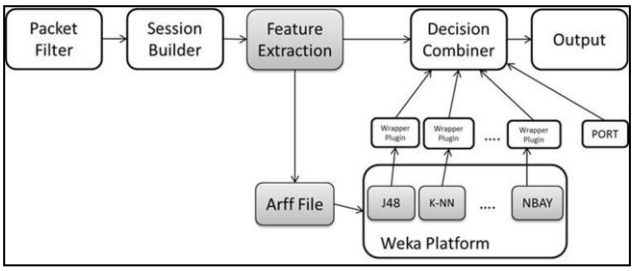


**Fig 4** Structure of the classification set up

## 5. Results

In our study, we split the data into three sets.

1. First set is used to train the base classifiers (20%).
2. Second set is used to train the gaiting network (20%).
3. Third set is used as a test set (60%).

We use four base classifiers. Three general purpose classifiers implemented in in WEKA: J48 Decision Tree(J48), K-Nearest Neighbor(KNN), and Naïve Bayes(NBAY). We also used a Port-based approach classifier (PORT) implemented in TIE. As a multi-classifier system, we used three decision combination strategies implemented in TIE: majority voting (MV), Naïve Bayes (NB) and Wernecke's method (WER). We compared the performance of those combiners to our proposed mix of experts (M.O.E) strategy. The results are in Table 5.1.

**Table 2** Decision combiners overall accuracy

| | **Base Classifier** | | | |
|---|---|---|---|---|
| | J48 | K-NN | NBAY | PORT |
| Overall accuracy | 93.3% | 87.6% | 39.7% | 18.7% |
| | **Decision Combiner** | | | |
| | MV | NB | WER | M.O.E |
| Overall accuracy | 90.4% | 42.3% | **94.7%** | 92.1% |

The best performing combiner was Wernecke's method, but our method has outperformed both Majority voting and Naïve Bayes. We believe that our method could have performed better than Wernecke's method with a different choice of base classifiers and input features. We plan to extend our experiment by trying different configuration of classifiers and features, and also inspect the effects of our proposed method on the early classification of flows.

## 6. Conclusion

In this work we have presented a new decision combination method for Internet traffic classification. The new method uses classifier selection instead of classifier fusion in order to enhance the overall accuracy of the base classifiers.

## References

[1] Kim, H., Claffy, K., Fomenkov, M., Barman, D., Faloutsos, M., Lee, K.: Inter- net traffic classification demystified: myths, caveats, and the best practices. In: CoNEXT 2008: Proceedings of the 2008 ACM CoNEXT Conference, pp. 1–12. ACM, New York (2008)

[2] Nguyen, T., & Armitage, G. (2008). A survey of techniques for internet traffic classification using machine learning. *IEEE Communications Surveys & Tutorials*, *10*(4), 56–76. doi:10.1109/SURV.2008.080406

[3] Bagui, S. C. (2005). *Combining Pattern Classifiers: Methods and Algorithms*. *Technometrics* (Vol. 47, pp. 517–518).

[4] A. Dainotti, W. de Donato, A. Pescapè, "TIE: a Community-Oriented Traffic Classification Platform", International Workshop on Traffic Monitoring and Analysis (TMA'09) @ IFIP Networking 2009 - May 2009, Aachen (Germany)

[5] Dainotti, A., Pescapé, A., & Sansone, C. , Early classification of network traffic through multi-classification. *Traffic Monitoring and Analysis*, 122–135, 2011.

[6] A. Moore and K. Papagiannaki. Toward the accurate identification of network applications. In PAM, April 2005.

[7] Dainotti, A., Pescapè, A., & Claffy, K. (2012). Issues and future directions in traffic classification. *Network, IEEE*, (February), 35–40.

[8] L7-filter, Application Layer Packet Classifier for Linux. http://l7-filter.sourceforge.net

[9] Tcpdump and the Libpcap library. http://www.tcpdump.org [November 2008].

[10] Internet Assigned Numbers Authority (IANA) website and port assignment www.**iana**.org/assignments/**port**-numbers

[11] *tcpanon* http://www.ing.unibs.it/ntw/tools/tcpanon/

[12] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.