Traffic Classification using a Statistical Approach

Denis Zuev[1] and Andrew W. Moore[2]

[1] University of Oxford, Mathematical Institute, zuev@maths.ox.ac.uk*
[2] University of Cambridge, Computer Laboratory, andrew.moore@cl.cam.ac.uk**

Abstract. Accurate traffic classification is the keystone of numerous network activities. Our work capitalises on hand-classified network data, used as input to a supervised Bayes estimator. We illustrate the high level of accuracy achieved with a supervised Naïve Bayes estimator; with the simplest estimator we are able to achieve better than 83% accuracy on both a per-byte and a per-packet basis.

1 Introduction

Traffic classification enables a variety of other applications and topics, including Quality of Service, security, monitoring, and intrusion-detection that are of use to researchers, accountants, network operators and end users. Capitalising on network traffic that had been previously hand-classified provides us with training and testing data-sets. We use a *supervised* Bayes algorithm to demonstrate an accuracy of better than 66% of flows and better than 83% for packets and bytes. Further, we require only the network protocol headers of unknown traffic for a successful classification stage.

While machine-learning has been used previously for network-traffic/flow classification e.g., [1], we consider our work to be the first that combines this technique with the use of accurate test and training data-sets.

2 Experiment

In order to perform analysis of data using the Naïve Bayes technique, appropriate input data is needed. To do this, we capitalised on trace-data described and categorised in [2]. This classified data was further reduced, and split into 10 equal time intervals each containing around 25,000–65,000 objects (flows). To evaluate the performance of the Naïve Bayes technique, each dataset was used as a training set in turn and evaluated against the remaining datasets, allowing computation of the average accuracy of classification.

Traffic categories Fundamental to classification work is the idea of classes of traffic. Throughout this work we use classes of traffic defined as a common group of user-centric applications. Other users of classification may have both simpler definitions,

* This work was completed when Denis Zuev was employed by Intel Research, Cambridge.
** Andrew Moore thanks the Intel Corporation for its generous support of his research fellowship

e.g., Normal versus Malicious, or more complex definitions, e.g., the identification of specific applications or specific TCP implementations.

Described further in [2], we consider the following categories of traffic: BULK (e.g., ftp), DATABASE (i.e., postgres, etc.), INTERACTIVE (ssh, telnet), MAIL (smtp, etc.), SEVICES (X11, dns), WWW, P2P (e.g., KaZaA, . . .), ATTACK (virus and worm attacks), GAMES (Half-Life, . . .), MULTIMEDIA (Windows Media Player, . . .).

Importantly, the characteristics of the traffic within each category are not necessarily unique. For example, the BULK category which is made up of ftp traffic, consists of both the control channel, which transfers data in both directions, and the data channel consisting a simplex flow of data for each object transferred. The assignment of categories to applications is an artificial grouping that further illustrates that such arbitrary clustering of only-minimally-related traffic-types is possible with our approach.

Objects and Discriminators Our central object for classification is the flow and for the work presented in this extended-abstract we have limited our definition of a flow to being a complete TCP flow — that is all the packets between two hosts for a specific tuple. We restrict ourselves to complete flows, those that start and end validly, e.g., with the first SYN, and the last FIN ACK.

As noted in Section 1, the application of a classification scheme requires the parameterisation of each object to be classified. Using these parameters, the classifier allocates an object to a class, due to their ability to allow discrimination between classes. We refer to these object-describing parameters as discriminators. In our research we have used 249 different discriminators to describe traffic flows, these include: flow duration statistics, TCP Port information, payload size statistics, fourier transform of the packet interarrival time discriminators — a complete list is given in [3].

3 Method

Machine Learned classification Here we briefly describe the machine learning (ML) approach we take, a trained Naïve Bayes classifier, along with a number of the refinements we use. We would direct interested readers to [4] for one of many surveys of all ML techniques.

Several methods exist for classifying data and all of them fall into two broad classes: deterministic and probabilistic classification. As the name suggests, deterministic classification assigns data points to one of a number of mutually-exclusive classes. This is done by considering some metric that defines the distance between data points and by defining the class boundaries. On the other hand, the probabilistic method classifies data by assigning it with probabilities of belonging to each class of interest.

We believe that probabilistic classification of Internet traffic, and our approach in particular, is more suitable given the need to be robust to measurement error, to allow for supervised training with pre-classified traffic, to be able to identify similar characteristics of flows after their probabilistic class assignment. We believe that the method be tractable and understood, and be able to cope with the unstable-dynamic nature of Internet traffic and that the method allow identification of when the model requires retraining. Additionally, the method needs to be available in a number of implementations.

Naïve Bayesian Classifier The main approach that is used in this work is the Naïve Bayes technique described in [5]. Consider a collection of flows $\mathbf{x} = (x_1, \ldots, x_n)$, where each flow x_i is described by m discriminators $\{d_1^{(i)}, \ldots, d_m^{(i)}\}$ that can take either numeric or discrete values. In the context of the Internet traffic, $d_j^{(i)}$ is a discriminant of flow x_i, for example it may represent the mean interarrival time of packets in the flow x_i. In this paper, flows x_i belong to exactly one of the mutually-exclusive classes described in Section 2. The supervised Bayesian classification problem deals with building a statistical model that describes each class based on some training data, and where each new flow y receives a probability of getting classified into a particular class according to the Bayes rule below,

$$p(c_j \mid y) = \frac{p(c_j) f(y \mid c_j)}{\sum\limits_{c_j} p(c_j) f(y \mid c_j)} \tag{1}$$

where $p(c_j)$ denotes the probability of obtaining class c_j independently of the observed data, $f(y \mid c_j)$ is the distribution function (or the probability of y given c_j) and the denominator acts as a normalising constant.

The Naïve Bayes technique that is considered in this paper assumes the independence of discriminators $d_1, \ldots, d_m$ as well as the simple Gaussian behaviour of them. The authors understand that these assumptions are not realistic in the context of the Internet traffic, but [5] suggest that this model sometimes outperforms certain more complex models.

4 Naïve Bayes Results

Our experiments have shown that the Naïve Bayes technique classified on average 66.71% of the traffic correctly. Table 1 demonstrates the classification accuracy of this techinique for each class. It can be seen from this table, that SERVICES and BULK are very well classified, with around 90% of correctly-predicted flows. In comparision to other results, it could be concluded that most discriminator distributions are well separated in the Euclinean space.

At this stage, it is important to note why certain classes performed very poorly. Classes such as GAMES and INTERACTIVE do not contain enough samples, therefore, Naïve Bayes training on these classes is not accurate or realistic. ATTACK flows were often confused with the WWW flows, due to the similarity in discriminators.

	WWW	MAIL	BULK	SERV	DB	INT	P2P	ATT	MMEDIA	GAMES
Accuracy (%)	65.97	56.85	89.26	91.19	20.20	22.83	45.59	58.08	59.45	1.39
Probability (%)	98.31	90.69	90.01	35.92	61.78	7.54	4.96	1.10	32.30	100.00

Table 1. Average accuracy of classification of Naïve Bayes technique by class and Probability that the predictive class is the real class.

Alongside accuracy we consider it important to define several other metrics describing the classification technique. Table 1 shows how traffic from different classes gets classified — clearly an important measure. However, if a network administrator were to use our tool they would be interested in finding out how much trust can be placed in the classification outcome. Table 1 also shows the average probability that the predicted flow class is in fact the real class, e.g., if flow x_i has been classified as WWW, a measure of trust gives us a probability that x_i is in reality WWW.

A further indication of how well the Naïve Bayes technique performs is to analyse the volume of accurately-classified bytes and packets. The results obtained are: 83.98% and 83.93% of packets and bytes, respectively, were correctly classified by the Naïve Bayes technique described above. In contrast port-based classification achieved an accuracy of 71.02% by packet and 69.07% by bytes (from [2]). Comparing results in this way highlights the significant improvement of our Naïve Bayes technique over the port-based classification alone.

5 Conclusions & Further Work

We demonstrate that, in its simplest form, our probabilistic-classification is capable of 67% accuracy per-flow or better than 83% accuracy both per-byte and per-packet. We maintain that access to a full-payload trace, the only definitive way to characterise network applications, will be limited due to technical and legal restrictions. We illustrate how data gathered without those restrictions may be used as training input for a statistical classifier which in turn can provide accurate, albeit estimated, classification of header-only trace data.

References

1. Anthony McGregor et al.: Flow Clustering Using Machine Learning Techniques. In: Proceedings of the Fifth Passive and Active Measurement Workshop (PAM 2004). (2004)
2. Moore, A.W., Papagiannaki, K.: Toward the accurate identification of network applications. In: Passive & Active Measurement Workshop 2005 (PAM2005), Boston, MA (2005)
3. Moore, A., Zuev, D.: Discriminators for use in flow-based classification. Technical report, Intel Research, Cambridge (2005)
4. Mitchell, T.: Machine Learning. McGraw Hill (1997)
5. Witten, I.H., Frank, E.: Data Mining. Morgan Kaufmann Publishers (2000)