Machine Learning in Internet Traffic Classification

From Csewiki

Contents

The introduction of IP Traffic Classification

The definition of IP traffic classification

The IP traffic classification is to infer the application-level usage patterns in time. Real-Time traffic classification has the potential to solve difficult network management problems for Internet Service Providers(ISPs) and their equipment vendors. It is useful for network operators to know what's flowing in the network promptly so that they can react quickly in support of various business goals. For example, internet traffic classification can be used to identify customer usage of network resources that in some way contravenes the operator's term of service. Another example is that traffic classification can be used in lawful interception under government's request.

Traffic classification metrics

a. Flow Accuracy: the accuracy with which flows are correctly classified, relative to the number of other flows in the author's test and/or training dataset(s). b. Byte Accuracy: How many bytes are carried by the packets of correctly classified flows, relative to the total number of bytes in the author's test and/or training dataset(s). note: Whether flow accuracy or byte accuracy is more important will generally depend on the classifier's intended use.

Limitations of package inspection for traffic classification

Traditional IP traffic classification relies on the inspection of a packet's TCP or UDP port numbers(port based classification), or the reconstruction of protocol signatures in its payload(payload based classification). Each approach suffers form a number of limitations. a. Port based IP traffic classification: 1. Some applications may not have their ports registered with IANA. 2. Sometimes, server ports are dynamically allocated as needed. 3. IP layer encryption may also obfuscate the TCP or UDP header, making it impossible to know the actual port numbers. b. Payload based IP traffic classification: 1. It imposes significant complexity and processing load on the traffic identification device. 2. This approach is difficult or impossible when dealing with proprietary protocols or encrypted traffic.

Classification based on statistical traffic properties

Newer approaches rely on traffic's statistical characteristics to identify the application. An assumption underlying such methods is that traffic at the network layer has statistical properties (such as the distribution of flow duration, flow idle time, packet inter-arrival time and packet lengths) that are unique for certain classes of applications and enable different source applications to be distinguished from each other. The need to deal with traffic patterns, large datasets and multi-dimensional spaces of flow and packet attributes is one of the reasons for the introduction of ML techniques in this field.

The application of ML in IP traffic classification

A class usually indicates IP traffic caused by (or belonging to) an application or group of applications. Instances are usually multiple packets belonging to the same flow. Features are typically numerical attributes calculated over multiple packets belonging to individual flows. Examples include mean packet lengths,

standard deviation of inter-packet arrival times, total flow lengths (in bytes and/or packets), Fourier transform of packet inter-arrival time, and so on. Since not all features are equally useful, so practical ML classifiers choose the smallest set of features that lead to efficient differentiation between members of a class and other traffic outside the class. The following three figures illustrate the steps involved in building a traffic classifier using a supervised learning algorithm.

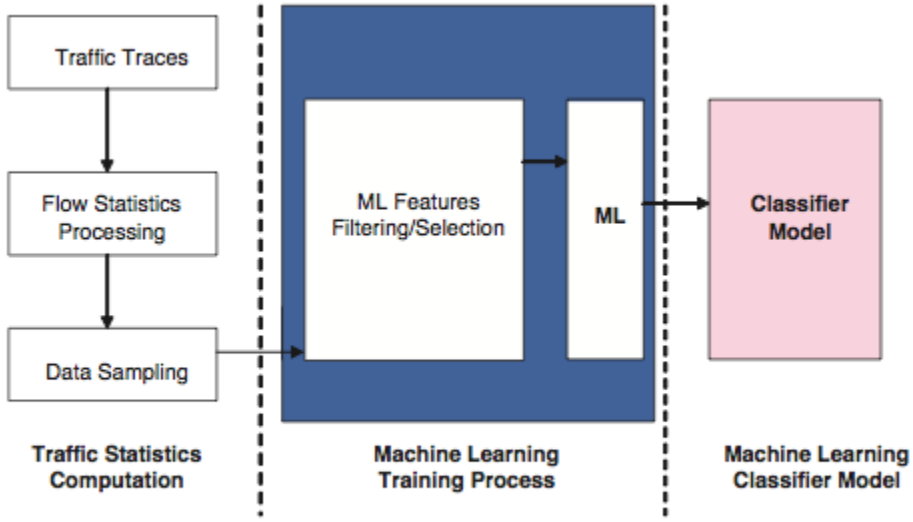

Fig. 3. Training the supervised ML traffic classifier

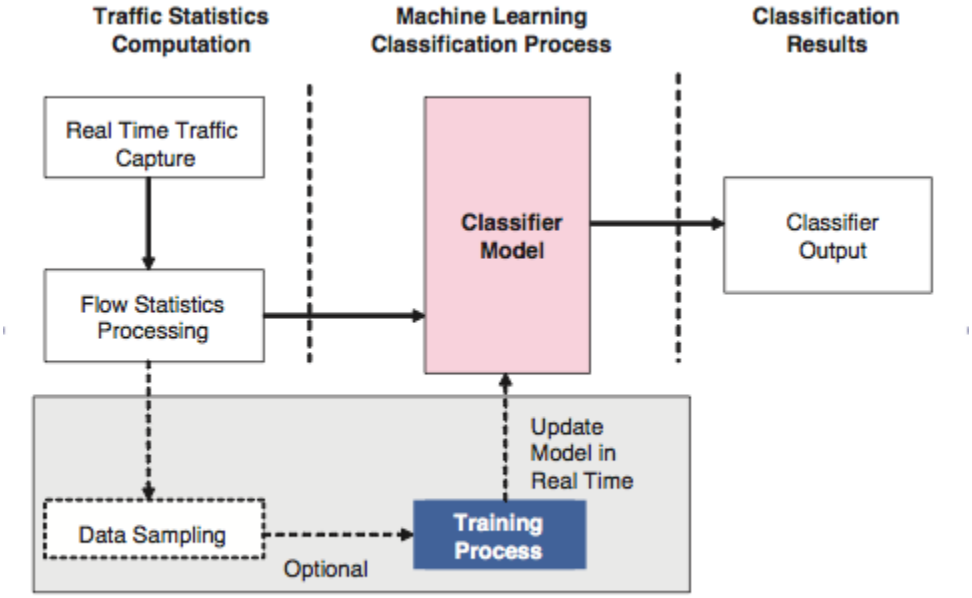

Fig. 4. Data flow within an operational supervised ML traffic classifier

Machine Learning based IP traffic

classification techniques

In the publications on ML-based IP traffic classification to date, there are three broad categories: Clustering approaches, supervised learning approaches and hybrid approaches. A brief introduction for each of them is given as follows:

Clustering Approaches

This category focuses on the employment of unsupervised machine learning techniques. One typical example is named as flow clustering using expectation maximization. This approach clusters traffic with similar observable properties into different application types.

Supervised Learning Approaches

This category focuses on the employment of supervised learning techniques. One typical example is to apply the supervised Machine Learning Naive Bayes techniques to categorize Internet traffic by application. Traffic flows in the dataset used are manually classified allowing accurate evaluation.

Hybrid Approaches

This category focuses on the combination of supervised and unsupervised learning techniques. A semi-supervised traffic classification approach, proposed by Erman in 2007, combine unsupervised and supervised methods. This approach has the following advantages: faster training time with small number of labeled flows mixed with a large number of unlabeled flows, being able to handle previously unseen applications and the variation of existing application's characteristics, and the possibility of enhancing the classifier's performance by adding unlabeled flows for iterative classifier training.

Conclusion

Currently, the usage of a number of different ML algorithms for offline analysis, such as AutoClass, Expectation Maximization, Decision Tree, Naive-Bayes etc. has demonstrated high accuracy (up to 99%) for a various range of Internet applications traffic. However, there is still a lot of room for further research in the field. For instance, each ML algorithm may perform differently toward different Internet applications, and may require different parameter configurations. The use of a combination of classification models is worth investigating. The application of ML algorithms for newer applications is still an interesting open

field.

Reference

Thuy T.T. Nguyen and Grenville Armitage, "A Survey of Techniques for Internet Traffic Classification using Machine Learning", 2008.http://caia.swin.edu.au /cv/garmitage/things/Nguyen_Armitage_SurveysAndTutorials2008.pdf