Wi-Fi encrypted traffic classification

Alessandro Zito, Carlos Santillan

August 2022

ID number:	890219 -
Personal code:	10617579 - 10659783
Supervisor:	Alessandro Enrico Cesare Redondi

Contents

1 Introduction

The goal of this project is to implement a Machine Learning classifier that can distinguish different types of traffic that a user performed by sniffing traffic in monitor mode. Monitor mode means that all the traffic that the Network Interface Card (NIC) can receive is captured, regardless of the destination.

2 Sniffing and dataset building

2.1 Sniffing

Sniffing and capturing the packets was the first step to building our dataset: the main problem was to capture it without using port numbers, IP addresses, and anything that goes beyond 802.11, cause the traffic in monitor mode is encrypted. First of all, we took the MAC address of the target device. Then with the help of *airmong-ng* [1] to put the NIC in monitor mode and using Wireshark as our packet sniffer software, we captured 20 minutes of each traffic type that we wanted to analyze, filtering packets destination and source with the MAC address chosen before

2.2 Dataset Building

After sniffing the packets, we had several .pcap files. To be able to analyze, we have to convert them The sniffing phase produced several .pcap files that required processing. To convert the capture files into a dataset ready for classification we had to extract meaningful features. We consider traffic flows lasting W seconds and computed:

- Mean packet length
- Variance of packet length
- Mean inter-arrival time
- Variance of inter-arrival time
- Maximum and minimum packet length
- Number of packets of type: QoS Data, QoS Null function, Other
- Number of packets sent and received

The choice of the parameter W was based on the ability to distinguish the traffic type in that amount of time. Starting from 5 seconds we chose 15 seconds since it yielded better results. So we processed the data to obtain the features and produced a dataset containing 477 rows

3 Data Exploration

We briefly looked at the variance of the features and noticed a significant unbalance that we addressed by normalizing the data. Since we had 11 features we performed principal component analysis to assess qualitatively the difference between the types of traffic. We kept the first 3 principal components (over 80% of variability explained) and plotted against each other. shows the pairs plot and we can see a pattern if we color the dataset by traffic type, they tend to appear in well-defined clusters. With this information, we proceeded to train a classifier using the original 11 features.

4 Classification and results

There were many options, ranging from KNN to Logistic Regression and Support Vector Machines. Since KNN didn't require any assumptions on normality and our data was well separated from the beginning, we selected it as our Machine Learning classifier. KNN classifies each sample based on its immediate neighbors so, for example, if my K-nearest neighbors are of type A, then I will be classified as A, or whatever the majority of my neighbors' type is. We used the Euclidean distance as the distance between the samples. We divided our dataset into training and test set (65% and 35% respectively) and to choose the hyper-parameter K, we performed leave-one-out cross-validation on the training set and the best K found was 4. So we tested the classifier against the test set and obtained an error rate of 7%

5 Conclusion

We believe that our work is coherent with the beginning requests; we consider the error rate output to be acceptable for this type of project. It may prove useful in monitoring network usage to know which type of traffic is used and improving the network resource allocation

Bibliography

[1] Airmon-ng. URL https://www.aircrack-ng.org/doku.php?id=airmon-ng.