CSC 480: Introduction to Data Mining

Fall 2018

Assignment 2: Data Mining for Cybersecurity

In this assignment, rather than using machine learning algorithms on nicely curated data sets such as those found in the UCI Repository, you will be dealing with real-world data. In particular, you will be using network traffic data generated by mobile apps on monitored smart phones collected by Dr. Zhen Liu and her research group. Dr. Liu and her group collected both active flows where users chose to share their data when using an app and passive flows where apps were launched on the phone and traffic collected while the phone was not used. The goal of your assignment is to solve two different problems:

1) Classify the active flow data according to the mobile app that generated it (e.g., QQ, WeChat, facebook, etc.).
2) Classify active from passive flows (Randomly sample a fixed number of examples (such as 20,000) from active data, and combine them with passive data for the passive flow detection experiment).

Mobile traffic classification is the foundation for QoS (Quality of Service) provision, bandwidth allocation and traffic shaping etc. For example, the high interactive applications (audio chat or video chat) require high QoS to avoid losing packets, so as to provide good user experience. Whereas some other applications (such as Browsers) do not need interactive communication, we could assign low QoS for the traffic generated by these applications

Mobile applications generate background traffic when the end-user is not actively using the app. If this background traffic could be accurately identified, network operators could de-prioritise this traffic and free up network bandwidth for priority network traffic.

The data was collected in the form of Network Flows. The data is available at: https://wangruoyu.github.io/mobilegt/. The data is described below:

Data description:

Active data files:

(1) biFeatureData: data are characterized by the bi-flow feature set

(2) uniFeatureData:data are characterized by the uni-flow feature set

They are characterized by different feature sets. You could try the algorithms on the two data sets and find out which feature set is better.

Passive data files

(1) Bipassivedata: data are characterized by the bi-flow feature set

(2) unipassivedata:data are characterized by the uni-flow feature set

In this assignment, your goal is to run various classifiers and try to combine feature-selection methods, class-imbalance approaches, outlier detection methods, and other such data filtering techniques that you see fit with algorithms such as Decision Trees, Neural Networks, Naïve Bayes, SVMs, k-NN, Bagging, Boosting, Random Forests, etc. to try to find a way to obtain good results on this data. Prior to running your experiments, take time to study your data set and see what kind of filters would be useful to apply to the data. Don't just try plenty of them. Instead, attempt to understand the data and reason about what approaches may be best given their characteristics. If you have any questions about the data, please contact our expert-in-residence, Zhen Liu at jeannylz@yahoo.com. She will be able to help you.