Machine Learning Research

Tim Birtles

This document summarises the research I carried out independent of my teammates to implement machine learning into the Network Traffic Profiler project.

ML Types

1. Supervised
    a. ML model is trained on labelled datasets
    b. Commonly used for image recognition, predictive analysis and fraud detection
    c. Comprises several algorithms:
        i. Regression – predict output values by identifying linear relationships been values. Focuses on one dependent variable and a series of changing variables. Includes linear regression, random forest.
        ii. Classification algorithms – predict categorical output such as "junk" vs "not junk", by drawing a conclusion from observations, trained on labelled input data. Includes logistic regression, k-nearest neighbours and support vector machines.
        iii. Naïve Bayes classifiers – Classification for large datasets. Includes decision trees which can work with regression and classification algorithms.
        iv. Neural networks – Simulating the human brain with linked processing nodes. Best for processes like language translation, image recognition, creation, and speech recognition.
        v. Random forest – Predict a value or category by combining the results from several decision trees.
2. Unsupervised
    a. Draw inferences from unlabelled datasets by using labelled datasets to train on.
    b. Facilitates exploratory data analysis, pattern recognition and predictive modelling.
    c. Allows for identification of associations between data in large databases, facilitating data visualisation.
    d. Algorithms:
        i. K-means clustering – assign data points into K groups, where the points closest to a given centroid are clustered under the same

category. Commonly used for image segmentation and compression

ii. Hierarchical clustering – describes a set of clustering techniques where data points are initially isolated into groups then merged based on similarity until one cluster remains.

iii. Probabilistic clustering – Groups datapoints on the likelihood they belong to a distribution.

3. Semi-supervised learning
    a. Trained on a small labelled and large unlabelled dataset, whereby the labelled data guides the process for the unlabelled data.

Python ML libraries:

- Scikit-learn – Built on NumPy and SciPy. Easy to use and supports several supervised and unsupervised algorithms.
- TensorFlow – Specialises in differentiable programming, Flexible and easy to develop.
- PyTorch – Typically for natural language processing. Fast at large data sets.

Problem and proposed solution

The use of ML within this project aims to classify user actions, such as a user 'liking' a YouTube video, versus 'disliking' or 'subscribing', therefore the use of a classification algorithm within a supervised learning environment is the best fit. The existing supply of training data from our client will aid in the process of training the AI. Scikit-learn seems to be the most suitable library for the project.

Asking AI how to proceed, See *AI Prompts (Semester 2).pdf* for full prompt and response.

- Raw PCAP data will be too noisy so features will need to be extracted from each flow such as:
    - Packet Length Stats (mean, median, standard deviation of packet sizes), as different actions will have different sized up/downstream bursts (think play vs comment).
    - Time between packets
    - Burstiness (Number of packets sent in quick succession)
    - Directional Ratio (bytes ratio for uplink vs downlink)

- Isolate flows directed towards the chosen Source (Youtube/Google).
- Split the test data into Training, Validation and Test sets.

References

https://www.ibm.com/think/topics/machine-learning-types

https://www.sas.com/en_gb/insights/articles/analytics/machine-learning-algorithms.html

https://www.coursera.org/articles/python-machine-learning-library