Course: EECS4480

Term: Summer 2021

Project Title: Feasibility of Network Traffic Classification Based on LDA and LSTM

Supervisor: name (Email): Dr. Sead Alrabaee (alrabaee@yorku.ca)

Student Name: Yuansen Zhu
Student Number: 215166895
Student's Email: larryzhu@my.yorku.ca

Code Output Report

The project is trying to determine the feasibility of classification network traffic with two kinds of Machine Learning Algorithms. By the research, learning and coding, I target a dataset that fits the project. And for the detail, I will be trying to build a model that analyses the connection periods with numbers of connections. Build an LSTM network that can predict the possible connection numbers at a particular time. With the feasibility of building a model that can calcsilicate the potential unlikely numbers of connections, LSTM also proves the feasibility of classifying the types of communications once using a proper dataset with specific factors (or indicators, such as numbers of connections this project) of network traffic. And analyse them for further use.

Introduction to data sets

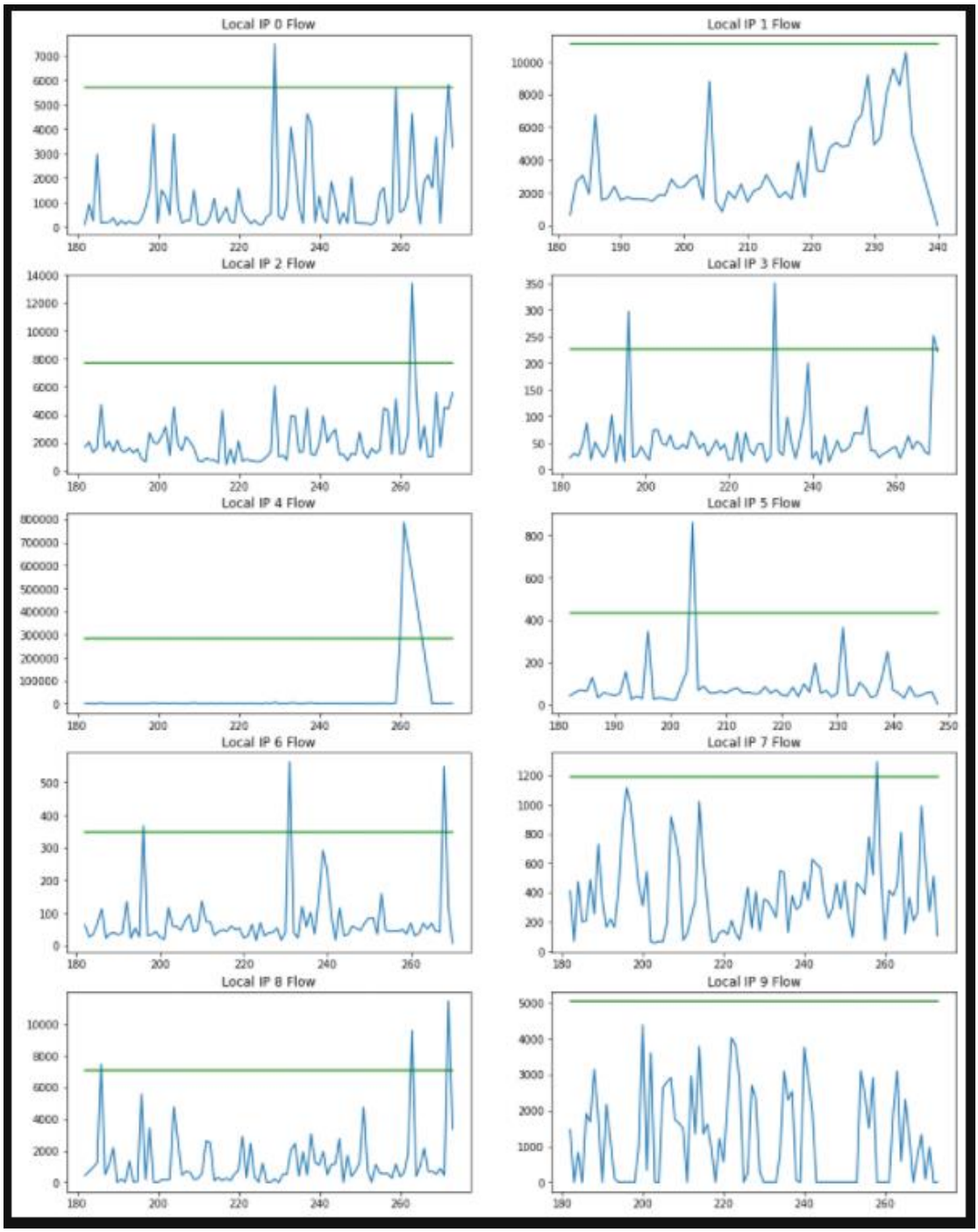
In this task, we use the following data sets: Computer Network Traffic Data -A ~500K CSV with summary of some real network traffic data from the past. The dataset has 21K rows and covers 10 local workstation IPs over a three months period. Each row consists of four columns:

- date: From 2006-07-01 through 2006-09-30
- l_ipn: local IP (coded as an integer from 0-9)
- r_asn: remote ASN (an integer which identifies the remote ISP)
- f: flows (count of connections for that day)

We will model based on this data set, use the historical information to predict the number of connections in the next stage every day, and visualize the output of the graph.

Data analysis

Firstly, the data is analyzed visually, and some useful information is obtained. First of all, we can visualize the connection times of each local IP. From the chart, we can see that the connection fluctuation of each IP flow is very large, but it also presents the regularity of wave shape, so it does not belong to the network data of random walk, which shows that our modeling has a good data base.



Data processing

Here we construct the training data set of the model, including the training data and the corresponding tags. First of all, the data sets are grouped and sorted according to the date and IP. Then the date field of the data is changed to the time type data, and some time attributes are extracted as the data features of the training set. In addition, because some fields of the data set have large values, it is easy to scale and train the model.

After data preprocessing, the feature data is extracted and saved. Finally, the dataset is segmented. Specifically, the historical data is used to roll back. The historical information of N days is extracted and fed to the model, and then the network connection of N + 1 day is predicted.

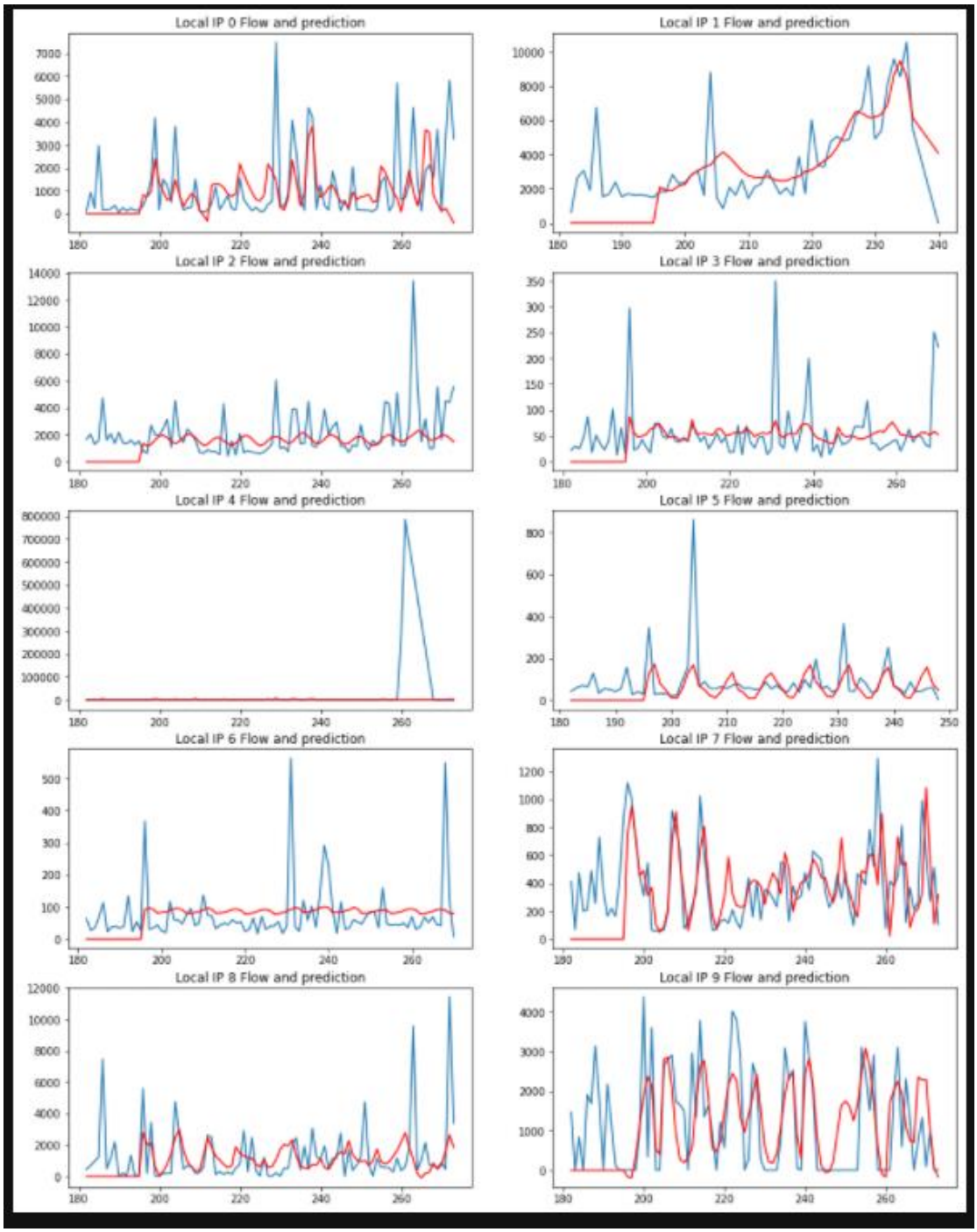
Model prediction

Because the data is time series data and contains historical information sorted by time series, LSTM recurrent neural network model will be used. Specifically, a two-layer recurrent neural network and a full connection layer are used to output the network; Some use super parameters as shown in the following table:

Hyperparameter	Value
Loss	Mean squared error
Optimizer	Adam
Epoch	200
Batch size	16

Results

The output results are analyzed visually, and the predicted value is compared with the real value (the red line is the predicted value, and the blue line is the real value). It can be seen that the performance of the model is not well fitted, and the predicted value does not show enough volatility.



To also prove the proper setting of models. The loose output for the model training history also shown the losses for each prediction. Therefore, the diagram showed the decreased losses of prediction value. As the input increase, the final proper prediction of the LSTM model is feasible for the network traffic classification.