MTRC train data analysis: A machine learning approach (2020)

Project Plan

Supervisor: Prof. Reynold CHENG
Team: Olivier KWOK (3035500839), Simon WONG (3035472355), Kalvin LEUNG (3035437939), Kayne CHUNG (3035453189)

Table of contents

1. Introduction

The railway traffic operation and schedule are very vulnerable and sensitive to delays and accidents. Over the past years, the MTR Corporation Limited (MTRC) is looking to utilise the historical track operation data collected to build a more intelligent modelling tool to estimate the railway traffic delay time, to enhance their resilience when handling these incidents.

Several studies are working on building tools for similar usage. However, take into account that the complexity of actual traffic conditions in every train station is quite different, for example, the characteristics of infrastructures and locomotives, number of interchanges, signalling system, railway design and platform capacity, the design of algorithm and parameters have to be adjusted and calibrated accordingly.

This project aims to develop a tool, driven by artificial intelligence (AI), that is capable of estimating the processing time with at least 95% of accuracy for any delays taking place along the Kwun Tong Line (KTL). With the aid of the predictions, MTRC will be able to make responses to delays and accidents more promptly and make a more appropriate judgement on their choices of executing the contingency plans.

This report serves as an introduction to the project plan. It consists of basic and theoretical backgrounds. Followed by the scope and objectives of the project, proposed methodologies and schedule. There is also a session listed several foreseeable risks and challenges ahead, and what kind of mitigations will be used to alleviate them.

2. Background

2.1 Mass Transit Railway (MTR)

MTR is one of the major public transport systems in Hong Kong. Comprising 10 heavy rail lines and a total of 95 stations, the network covers Hong Kong Island, Kowloon, and the New Territories with a ridership of over 5 million on a typical weekday [1]. At peak hours, train intervals can go as low as 2 minutes [2], making it one of the busiest metro systems in the world.

2.2 Problem that MTR faces

Albeit boasting a 99.9% on-time passenger service [1], the MTR operates at almost full capacity at peak hours and occasionally suffers from delays due to various reasons such as signalling faults, technical failures and overcrowding. In the past decade, there have been over 200 short delays (8-30 minutes) and around 14-20 long delays (>30 minutes) each year [3]. These disruptions to train service not only cause serious inconvenience to MTR commuters but also add a significant burden to road traffic in a city heavily relying on the metro system. As such, reducing train delays has been identified as a key area of improvement by MTRC in ensuring mobility in the city.

2.3 Modeling the metro network

Beyond the Hong Kong MTR network, many large urban railway systems, such as the metro in Beijing, London and New York are facing the problem of overloading capacity [7]. Solutions to increase efficiency and service quality therefore are getting attention in the intelligent transportation community [4],[5],[6],[8]. One of the directions on solving this problem is to model the metro system such that accurately predicting the ridership and commute time is possible by applying various machine learning techniques [4],[6].

a. Statistical learning

Statistical learning is a framework for machine learning drawing from the fields of statistics. Statistical learning algorithms can build a predictive function based on certain factors from historical data. In the case of railway system analysis, a combination of weather records, historical train delay records and train schedule data is proven to be the most important factors influencing train delays in general case [5]. Previous studies applied various statistical learning algorithms such as linear regression and regression trees on large scale railway systems in China and the Netherlands [4],[5], the results indicated that train movement can be accurately predicted with sufficient historical data to support the learning model. However, factors determining train movement could be different in different railway systems, it is important to

fine-tune the learning model in this project to improve the accuracy of prediction and utilise the historical track operation data.

b. Deep learning

Deep learning is an important branch of the field of AI. Deep learning algorithms enable computers to autonomously learn from data and perform computations to give meaningful information, without being explicitly instructed by human-designed programs. A previous paper [6] demonstrated that using a graph convolutional network can predict the ridership of every station by modelling the metro network into a graph. Another research also indicated that a network loading model can be created to predict journey time and station passenger flow [8]. It also shows the relationship between can be drawn by analyzing the data collected from automated fare collection and automated vehicle location [8]. Researches have shown with proper modelling and suitable dataset, a neural network model could be built to leverage the historical data to forecast journey time and further predict the delay when an incident happens.

2.4 Previous work

This project has been running for more than 12 months and this year's goal is to build on top of last year students' work, improve the accuracy of the prediction results by adapting the machine learning approaches and to consolidate all the test results generated so far to draw a comprehensive analysis of the effectiveness of our AI model.

Last year only a very limited amount of data was provided by MTRC. This year, at least 2 months of train schedule data files are given and also log files regarding passengers flow information will be furnished by the other research groups. It is expected that some cross-comparison will be made between these different types of data at the later stage of the project.

3. Scope and Objectives

The main goal of this project is to work with MTRC, to develop a software platform for train movement prediction which includes the amount of time the train stays at the platform, the departure time of the train, and the amount of time the train stops in a tunnel. The project aims to make accurate predictions on train movement in a machine learning approach such that train operators can reduce the number of trains that suffer from a delay during accidents. The ultimate objective of the project is to minimise the impact of a disruption and hence improve the customer experience in MTR ride.

Multiple machine learning algorithms for example linear regression, regression trees, random forest and graph convolutional networks will be evaluated and merged to build a final deliverable. The training set of the machine learning models is limited to the trains log and passenger flow data provided by MTRC. Other variables such as weather conditions or public holidays would not be considered in the training process. For the trains log data, the project is specifically focused on KTL in particular and other lines of service would not be taken into account until the completion of the KTL model.

4. Proposed methodologies

The following section will discuss the planned methodology of this project and the set of technology plans to use with the explanation of the rationale behind. The project will proceed in a 3-phased approach, preparation, exploration and application. The preparation phase will focus on the review of previous work and data collection. Then, with the knowledge we consolidated in the preparation stage, different machine learning models will be developed and experimented in the exploration phase. In the application phase, a final machine learning application for MTR train delay prediction will be created with the best-performing model in the exploration phase.

4.1 Preparation

The first phase of the project will be the preparation work. The goal in this phase is to understand the previous work from last year and the raw data received from MTRC.

Since this project is a continuation of last year's project, the previous team may already have done some work, such as data cleaning and the test may be useful for this project. Therefore, it is beneficial to understand their work and learn from their experience before we conduct our study. We will review the code repository and reports from the previous project, consolidate their finding and extract tools or insight that will be useful to our project.

4.2 Exploration

The second phase will be the exploration of different machine learning methods with the data we prepare in phase one. Our team identified two vastly different states of art machine learning methods during the literature review, i.) predictive model of train running time and dwell time [4] and, ii.) passenger flow prediction by graph convolutional network (GCN) [6]. They both show good performance in modelling the metro system in recent research. Therefore, we plan to experiment with both frameworks simultaneously. Our goal in this phase is to test the effectiveness of these two networks and opt for the best one for application in the final phase.

4.2.1 Predictive model of train running time and dwell time

From the literature review, we found that delay time can be modelled by applying various statistical learning techniques on the train occupation data of a railway system [4]. Thus, we plan to implement a similar model with train data of KTL with statistical learning techniques: linear regression, random forests and regression trees.

a) Data preparation

We plan to make use of the raw operating data of KTL provided by MTRC. These data are the logs of historical data of the actual train movement in KTL. The train movement duration and schedule will be extracted and preprocessed from these raw logs to form a dataset. We plan to

use Python with Numpy and Pandas libraries for the data transformation job and then store the dataset in CSV format.

b) Implementation

Various statistical learning models, namely linear regression, random forests and regression trees will be used to model the train occupation data. We plan to implement these models in Python with library sci-kit learn and train them on the preprocessed dataset we prepared. Then measure the accuracy of each learning model and pick the best performed one.

4.2.2 Passenger flow prediction by graph convolutional network (GCN)

The passenger flow prediction is a model of the metro system from the perspective of passenger commute time. It predicts the passenger commute time by the relationship between stations using a graph convolutional network [6]. We plan to create a graph convolutional network with the reference to previous research.

a) Data collection

The commuter data of passengers will be used in the model since they provide the entry/exit station as well as the timestamp of their action which can be used for calculating the journey time. The commuter data can be extracted from the transaction logs of every ticket barrier in the MTR network. To preserve anonymity, customer identification of every record is cleaned and randomized. Similar to another modelling problem, Python will be used to transform the data into a dataset and store in CSV format.

b) Implementation

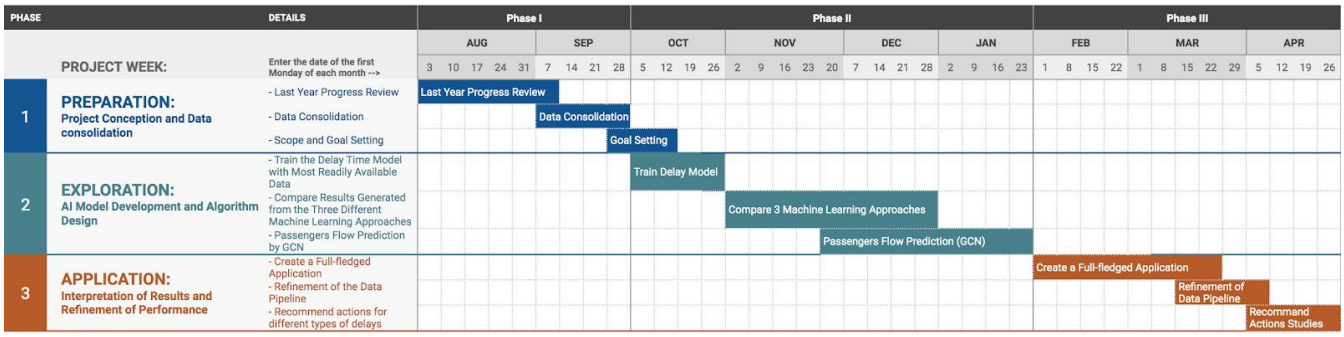
With the reference of previous research, the network will be trained by three graphs, a physical graph, a similarity graph and a correlational graph [6]. The physical graph will be the graphical representation of the train network, where edges represent the physical connection between stations and vertices are stations. The similarity graph is the passenger flow similarities of stations, where edges represent the similarity between two stations. To reduce complexity and dimensions, only top-k important edges will be used where k is going to be determined by experiments. The correlation graph is the representation of the correlation of passenger flow between stations. Similar to the similarity graph, the correlation graph is also constructed by using top-k important edges. The neural network will be trained by using these graphs with the actual passenger flow recorded in every 15 mins. It will be coded by Python with the use of Pytorch, one of the states of art Machine learning libraries. To accelerate the training progress, the training procedure will be deployed to HKU GPU farm to utilize GPU acceleration.

4.3 Application

The third and the last stage will be the application of the machine learning method we prototyped in the exploration stage. A fine-tuned final model will be created in this phase. Together with the prediction system, a data pipeline will be created. The goal at this stage is to

create a full-fledged application for MTR company to use this prediction model in a production environment.

A “Gantt Chart” is shown below, outlining the project timeline and detailed schedule for each phase.



5. Risks, Challenges & Mitigation

5.1 Unsatisfactory results from proposed methodologies

Although the proposed machine learning models show promising prediction accuracies in the experiments detailed in their respective research papers, it is possible that they do not perform as well on the MTR network as their general applicability on metro systems is not yet established. For instance, the predictive models discussed in section 4.2.1 were tested on a relatively small railway section [4], in contrast to the full-scale metro system this project concerns. Prediction performance of the models could vary due to differences in a range of attributes such as track characteristics, train schedule, ridership patterns, station topology etc. This is the reason why we plan to explore two different approaches simultaneously such that we could select the better one at a later stage. We may also further identify a few other methods in literature as backups.

5.2 Availability of data from MTRC

The relevant data provided by MTRC are what makes this project possible. While MTRC has pledged to support this project by giving us access to whatever types of data we need as far as possible, certain types of data may not be available to us due to technical constraints, or business and privacy concerns. Since this would limit the analysis methods we could use, we have to be in close communication with MTRC and make our requests at an earlier stage. Besides, it may also take time for them to extract and compile the data as per our requests. So to keep our project on schedule, it is important to let MTRC know our needs on time once we have decided certain data are necessary for the project.

5.3 Security of MTRC data

The data we use is confidential corporate information provided by MTRC solely for the agreed objective of this project and any form of disclosure of the data to third-party would severely undermine our collaboration with MTRC and hence the success of the project. Therefore, we are obliged to take data security seriously. All data provided by MTRC will be stored and processed only on the personal computers of the team members and the locations designated by the supervisor.

References

[1] Mass Transit Railway Corporation, “Business Overview,” 2020. [Online]. Available: http://www.mtr.com.hk/archive/corporate/en/publications/images/business_overview_e.pdf. [Accessed: Oct. 3, 2020].

[2] Mass Transit Railway Corporation, Service Hours, n.d. [Online]. Available: http://www.mtr.com.hk/en/customer/services/train_service_index.html. [Accessed: Oct. 3, 2020].

[3] Research Office of the Legislative Council Secretariat, MTR train service performance, 2017. [Online]. Available: https://www.legco.gov.hk/research-publications/english/1718issh07-mtr-train-service-performance-20171220-e.pdf. [Accessed: Oct. 3, 2020].

[4] P. Kecman, and R. M. P. Goverde, "Predictive modelling of running and dwell times in railway traffic." Public Transport, vol. 7, no. 3, pp. 295-319, 2015

[5] P. Wang and Q. Zhang, “Train delay analysis and prediction based on big data fusion.” Transportation Safety and Environment, vol. 1, no. 1, pp. 79-88, 2019

[6] L. Liu, J. Chen, H. Wu, J. Zhen, G. Li, and L. Lin, “Physical-Virtual Collaboration Modeling for Intra-and Inter-Station Metro Ridership Prediction”, arXiv e-prints, 2020.

[7] Koutsopoulos, H. N., Z. Ma, P. Noursalehi, and Y. Zhu, Chapter 10 - Transit Data Analytics for Planning, Monitoring, Control, and Information. In Mobility Patterns, Big Data and Transport Analytics (C. Antoniou, L. Dimitriou, and F. Pereira, eds.), Elsevier, 2019, pp. 229 – 261.

[8] Mo, Baichuan & Ma, Zhenliang & Koutsopoulos, Haris & Zhao, Jinhua. (2020). Capacity-Constrained Network Performance Model for Urban Rail Systems. Transportation Research Record: Journal of the Transportation Research Board.