HIGHWAY TOLLGATES TRAFFIC FLOW PREDICTION USING MACHINE LEARNING ALGORITHMS

Dai Yirui, Dong Meirong, Lim Chong Seng Hermann

Institute of Systems Science, National University of Singapore, Singapore 119615

ABSTRACT

Along with the fast development of interconnected economy, modern-day traffic networks have become larger and more complex. Ineffective management of these traffic networks have multiple negative impacts on the individual as well as the national economy.

The study aims to use machine learning algorithms such as XGBoost and LightGBM to predict the conditions at traffic bottlenecks such as highway tollgates to reduce congestion, enhance road users' experience and improve infrastructure management.

There are two main issues focused in this study, the first is to forecast the average traveling time through the tollgates and the second is to estimate the traffic volume at the tollgates.

Index Terms— One, two, three, four, five

1. INTRODUCTION

Traffic management is an increasingly difficult task for almost every government and traffic authorities due to rapid globalisation and greater population mobility. Poor traffic management can adversely impact quality of life and economic productivity due to rising fuel consumption, cost of movement and number of accidents. Although the consequences are alarming, traffic congestion remains a prevalent issue in many countries because they have inadequate resources to address the problem.Therefore, minimising traffic problems with optimized costs while satisfying the growing demand for more efficient traffic networks should be the core objective of today's traffic management studies.

In this study, the team aims to use machine learning algorithms such as XGBoost and Neural Network to predict traffic flow volume pattern and travel time patterns would occur in the near future(e.g.next 2 hours) on routes to help traffic authorities and commuters to make informed travel decisions, to divert traffic and to address the core objective.

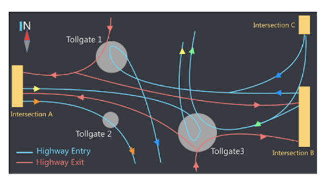
The study uses Knowledge discovery in databases (KDD) 2017 dataset for analysis. The dataset is provided by "Hanzhou Jiaotong Amap", the task is to predict highway tollgates traffic flow and the results are to be measured and ranked based on the "Mean Absolute Percentage Error (MAPE)" score.

The dataset is chosen because manual tollgates are usually the bottlenecks of traffic network and can easily cause traffic congestion and overwhelm traffic authorities and commuters. One small improvement made to the bottlenecks could easily be one giant leap for traffic conditions across the route, by Pareto principle.

- The main report is provided in *.tex file
- The reference is provided in *.bib file
- The figures are provided as separate jpg/png files

2. PROBLEM OVERVIEW

The road network topology (as shown in Fig.1) is a directed graph formed by a sequence of interconnected road links. There are 6 routes in the network.



intersection_id	tollgate_id	link_seq
A	2	110,123,107,108,120,117
A	3	110,123,107,108,119,114,118,122
B	1	105,100,111,103,116,101,121,106,113
B	3	105,100,111,103,122
C	1	115,102,109,104,112,111,103,116,101,121,106,113
C	3	115,102,109,104,112,111,103,122

Fig. 1. Visualization of Road Network Topology

Problem 1 is to estimate for every 20-minute time window the average travel time of vehicles for a specific route:

- Routes from Intersection_id A to Tollgate_ids 2 3;
- Routes from Intersection_id B to Tollgate_ids 1 3;
- Routes from Intersection_id C to Tollgate_ids 1 3.

Problem 2 is to predict average tollgate traffic volume for every 20-minute time interval window where tollgates 1 and 3 has entry and exit traffic and 2 has only entry traffic.

There are 5 classes of data provided in the dataset which are links, routes, trajectories, volume and weather.

Mean Absolute Percentage Error (MAPE) is used to evaluate the accuracy.

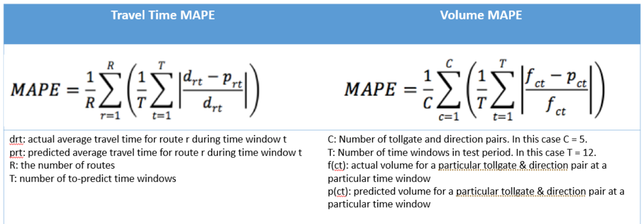


Fig. 2. Visualization of Road Network Topology

3. RELATED WORK

The existing works on KDD 2017 have proposed many solutions for highway tollgates traffic prediction. Some of which are XGboost, Radial Basis Function (RBF) nets and deep neural networks. The most favourable model is XGboost. Therefore, the team uses tuned XGboost models (Chen, 2019) as the main tool to generalize the regression model in this study.

Remarkably, XGboost has received tremendous attention in recent years (Morde Setty, 2019) because of its promising speed and performance (Brownlee, 2016). The top few winning teams of KDD cup 2017 competition have also recommended this learning algorithm for their predictions.

An alternative approach is to use a set of continuous distributions to approximate the regression via RBF. By using RBF, the idea is to combine (or sum) a set (kernels) of curves (Gaussian distributions) with various control parameters (std mean) to generalize the regression model (Deshpande, 2017).

More complex approaches that can stack and vote from different model (LASSO, GBDT, ADABoost and RandomForest etc) layers and deep neural networks such as(CNN-LSTM-Attn) and Temporal-Spatial-LSTM (TSLSTM) are also examined. However, details including how the models are implemented and their parameters are not provided by the author. This complex approach managed to reach a MAPE value around 0.104 for volume prediction.

In the end, due to time constraints, the team decided to follow the proposal submitted by the 1st place team in KDD cup 2017 competition to implement XGboost.[1]

4. PROPOSED APPROACH

With reference from KDD cup 2017 winning team's proposal, the team has proposed following step by step approach:

Phase	**Action taken**	**Action Category**
Study-1	Aggregate raw trajectories data based on 20-mins time window	Coding
Study-2	Plot volume of all time window at a particular day (1st day)	Exploring
Feature-1	Process and select feature from link and route tables for data training and prediction (link_counts, route_length, double_in_link_count, double_out_link_count, min_width)	Exploring
Feature-2	Process the features from trajectories table and clean outliers (add time lag for 2 hours, because based on other research works, the average traveling time from previous 2 hours have more significant impacts on the data)	Exploring
Feature-3	Process and understand the data from weather table	Exploring
Feature-4	Combine all the tables	Exploring
Base-1	Implement baseline approach	Coding
Improve-1	Test the baseline with different features and parameters to improve accuracy	Experimenting

Average Travel Time Prediction

Average Volume Prediction

Phase	Action taken	Action Category
Study-1	Aggregate raw volume data based on time window (20 mins)	Coding
Study-2	Plot volume of all time window at a particular day (1st day)	Exploring
Study-3	Look for irregularities and try to explain the reason. Exclude irregularities if not helpful for prediction	Exploring
Study-4	Plot volume by day for the whole training period	Exploring
Study-5	Look for irregularities and try to explain the reason. Exclude irregularities if not helpful for prediction	Exploring
Study-5	Look for irregularities and try to explain the reason. Exclude irregularities if not helpful for prediction	Exploring
Study-5	Look for irregularities and try to explain the reason. Exclude irregularities if not helpful for prediction	Exploring
Study-6	Plot top (4) frequent ‘travellers’ and see if there are similar travelling patterns (Explain for peak and trough)	Exploring
Base-1	Implement baseline approach	Coding
Base-2	Experiment with denoised data (excluded holiday periods), calculate baseline MAPE as benchmark	Experimenting
Base-3	Parse and join weather data with volume data	Coding
Base-4	Further experiment with weather data, check if MAPE improve	Experimenting
Improve-1	Aggregate raw volume data based on time window (20 mins) but include ‘has$_etc'asfeature$	Coding
Improve-2	Implement function to calculate mean volume values of each time window for past n (5) days	Coding
Improve-3	Exclude irregular high volumes at midnights	Experimenting
Improve-4	Experiment with volume vs. log(volume)	Experimenting
Improve-5	Experiment with different evaluation metrics: ‘mae’	Experimenting

5. EXPERIMENTAL RESULTS

$$B_{r,c} = \sum \{f(i,j) | (i,j) \in \Omega_{r,c}\}. \tag{1}$$

$$\sum_x = a + b + \hat{c}, \tag{2}$$

An inline equation is $a + b = c$. An example of two-column figure is provided in Figure 3, and the single-column figures is provided in Figure 4.

Table 1. The performance comparison.

Approach	Ref. [1]	Ref. [1]	Proposed approach
Metric A	0.8181	0.9171	0.9616
Metric B	0.8236	0.7654	0.8615

6. CONCLUSIONS

Summarize your key results. What are limitations of your approach? Suggest ideas for future extensions of your ideas.

7. REFERENCES

[1] D. Adams, *The Hitchhiker’s Guide to the Galaxy*, San Val, 1995.

Fig. 3. Test figure (two-column).

(a) (b) (c)

Fig. 4. Test figure (single column).