二項分類器による乗客数の予測
Predicting Passenger Number via Binary Classifier

学籍番号　47186834
氏名　陳　柏嘉 (Chen, PoChia)
指導教員　柴崎　亮介

1 Research background

People living in Tokyo mainly depends on public transportation,like the train, to satisfy their traveling need, which increases the potential risks as well. In the communication time, it easily gets crowded and chaos at the platform and stations. At that moment, the risk grows up. Hence, there is urgent need for those facilities and departments to get known the possible situation before the risk comes out as early as possible, so that they can prepare and prevent any miserable mistake.
Fortunately, with the rapid development of location acquisition technology and upcoming 5G mobile technology, we have multiple methods to record, analyze and capture the human mobility pattern. This will highly benefit the transportation scheduling, urban regulation, and even emergency management.
In academic field, there have been already a lot of superior researches about mobility prediction, however the topic about long-term prediction based on long-term historical trajectory is still on the way. For long-term future prediction, the most essential challenge is how to make the model understand and capture the features in the extremely long-term trajectory. And it is even harder to reproduce the individual mobility patterns. We believe the the long-term future prediction is more meaningful than those shot-term future prediction because it has the ability to give a more holistic view of human mobility.
In the field of ubiquitous computing, human mobility analysis and prediction have been comprehensively explored from individual scale to city wide scale, while most of them are focusing on multi-class predictor in stead of a binary predictor. However their accuracy cannot satisfy the reality need because the multi-class predictor provides too many options resulting that none of the class option can have a really accuracy. Besides classification, regression is one of the idea to tackle the mobility prediction problem either, which can give an exact location in the future. In our case, we don't care or consider the exact location of a person, so we tend to solve this as a classification problem.

2 Research purpose

In order to solve the long-term prediction problem we mentioned in the previous section, we are particularly interested in using long-history trajectory and meta information to predict the human mobility with a high accuracy and in higher resolution (1 km^2) via binary classifier.
In this study, we propose a framework including a complete data-processing procedure and a machine learning architecture based binary predictor.
First, we process the GPS trajectory with a 5-minute interval, so there are 288 positions in a day. Secondly, we turn the latitude and longitude into grid which is a method to represent the location for a specific area, which can reduce the computing burden. Third, before the training, we will balance the number of positive and negative samples in each epoch of dataset. Then, we train and evaluate the model with our dataset. Eventually, we predict whether a person will enter the station at a time of next week. And we also explore the difference between different models with our MetaInfo method through several experiments.
For the best result of our research, it is possible to give a long-term prediction of the passenger number with high accuracy and in higher resolution by analyzing the long-term historical trajectory. And the length of historical trajectory can be adjusted according to the need. Eventually, we can build up one binary predictor for each station, and assemble the results together to give a general prediction in the city scale. With the higher accuracy result, the user can execute relevant measure to achieve better services.

3 Methodologies

Fig. 1 shows the framework of my research. In the data-processing procedure, there are 4 steps:

- Interpolate the raw trajectory into trajectory with a 5-minute interval.
- The origin data are points-like data. We transfer the trajectory data into grids to represent location, because there is no necessity to predict the precise location as latitude and longitude. What we are concerned is the whole situation of the area which in our experiment is 1 km^2.
- Existence labeling: Using trajectory data over a period of time in the future such as 1 month, we add binary labels to represent the existence of a person at a specific time and area.
- Extract the processed trajectory with specific interval to reduce unnecessary computing.

Then, we produce the meta information manually. The meta information can include the multiple information of next week, like datetime information, event information, holiday information and etc. To build up the meta information of a week, we utilize a vector to record the meta information of next week, e.g. datetime information-(2,8) which means 9 o'clock on Wednesday morning. If we have more meta information of next week, we can encode the information and add into the vector. After this preparation, we do the oversampling operation on trajectory data and meta data pairs to generate our training dataset.

In the deep learning architecture, we are building up our models based on several neural network structures, testing the performance on long-term trajectory prediction task with those models, and eventually choose the well performance models out and compare with the typical time series statistic model. In this part, we are trying to embed the meta information of a week with historical trajectory of last week to capture the regularity of mobility, aiming to dig out the correlation between historical trajectory and meta information In *Methodology* chapter, we give not only the overview of our framework but also the detail

Through this method, it is possible to create a binary classifier for a week and even a longer time period. And in chapter *experiment*, we compare the performance of different models and do further performance test in case study to compare the machine learning methods with typical statistic models.

4 Experiment results

In chapter *experiment*, we first explore the model variants, including Hidden Markov model, RNN, LSTM, GRU, GLU and LightGBM. Except the performance of those models themselves, we also do the research about how our method help improving their performance on long-term prediction task. The results show that GLU+MetaInfo and LightGBM+MetaInfo have the superior performance and great efficiency when doing training and prediction. Second, we explore the influence of different trajectory length as input. We found that though the longer trajectory does help improve the performance of prediction, the longer input also causes lower efficiency, which means the training and prediction time increases. And we think it is not worthy. In section *Model variants* and *Hyper-parameter*, there are tables showing the model performance on test dataset, whose data are sampled from a complete dataset. Eventually considering the speed and accuracy demand, we do the case study with ARIMA, SARIMA, GLU+MetaInfo and LightGBM+MetaInfo, and analyze their different performance. Since the figures are too large and the number of them are too much, we left the figures in *case study* section. Parts of the results about case studies is in Fig.2 and Fig.3.

5 Conclusion and future work

In this research, we proposed a binary classifier idea, aiming to predict the passenger number in the station by predicting whether a person will enter the station or not. In the experiment part, we conduct series of preprocess procedure to enhance the model performance. And we evaluate the machine model performance with different experiment settings. Eventually, considering the accuracy and efficiency we select ARIMA, SARIMA,

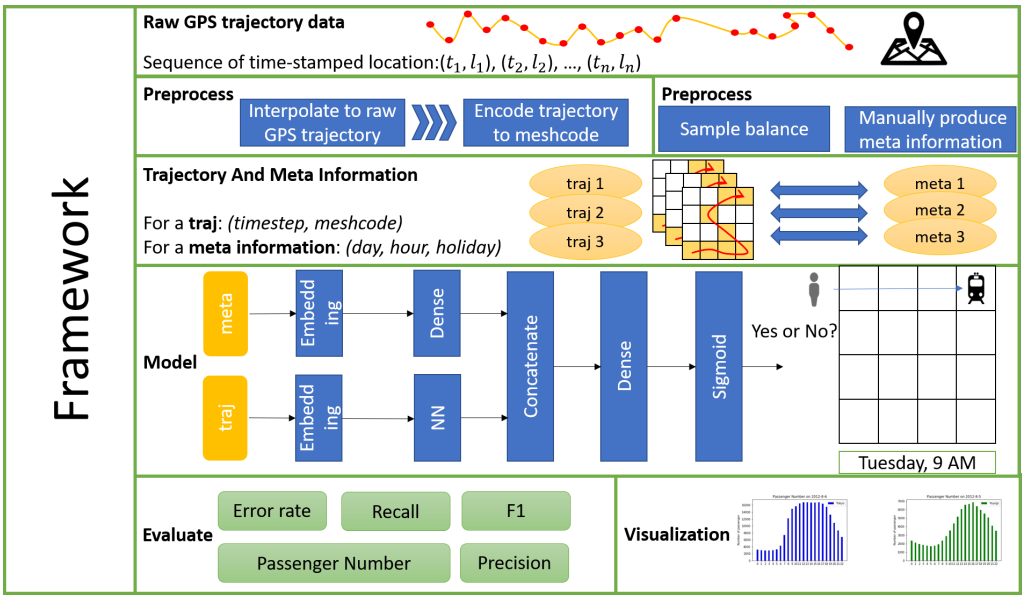


Fig. 1: Overview of our method

GLU+MetaInfo and LightGBM+MetaInfo to conduct the case studies at Tokyo station during different time periods. According to the three case studies, we have the following conclusions:

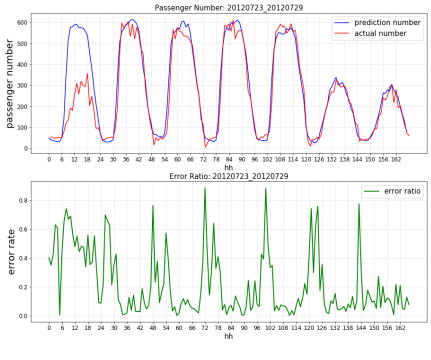
1. When facing the situation that the passenger trend appears under a specific and stable pattern, the statistic methods perform more stable than machine learning methods.

2. When the passenger trend appears different comparing with historical data the machine learning methods perform better than statistic methods.

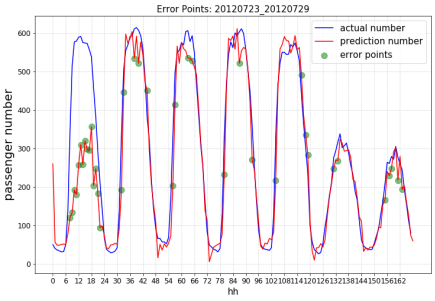
3. All of the methods possess the same problem: the prediction results follows the historical data slightly when the passenger trend is under a specific and stable pattern.

4. The LighGBM+MetaInfo and ARIMA model performs best in the case studies.

In summary, we believe that the binary classifier can capture the features in the long-term mobility pattern but there is still work can be done to improve the performance.
For the purpose of making the long-term mobility more reliable, there could be several future works for improving the performance of binary classifier based method. First, more feature engineering could be done to explore the internal correlation of regular mobility pattern because a lot of data features are lost when we transform the trajectory into mesh code format. Secondly, there is still improvement space in the preprocess procedure, e.g. spatiotemporal points extraction. At the end of current research, we found that our preprocess method missed some potential passengers, leading to the information lost.
Thirdly, in the previous data exploration, we found that the positive and negative sample ratio is extremely low. To figure the imbalanced data problem, we utilize sample balance method to ensure the model learns the positive and negative sample equally. However, even with repeat training, there are still unknown negative samples existing. To tackle this extreme imbalance situation, there could be researches regarding the positive samples as abnormal samples, and make the prediction by finding out the possible abnormal case in the long-term mobility pattern.

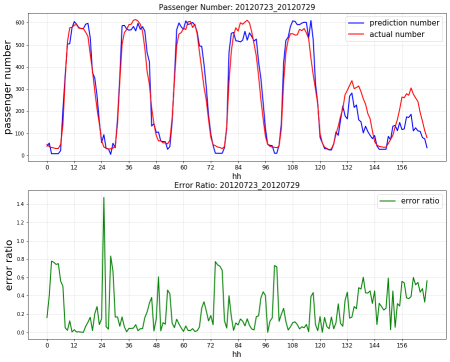


(a) Passenger number and error ratio

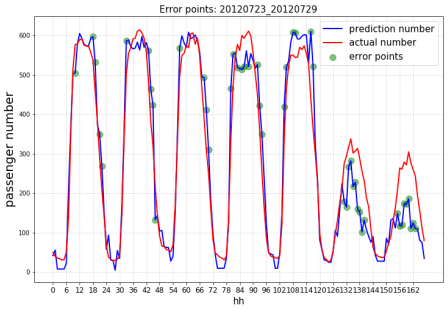


(b) Specific error points: when passenger number is over 200 and error ratio is over 10%

Fig. 2: Case 1: Predicted passenger number, error ratio, and specific error points of a week from ARIMA model



(a) Passenger number and error ratio



(b) Specific error points: when passenger number is over 200 and error ratio is over 10%

Fig. 3: Case 1: Predicted passenger number, error ratio, and specific error points of a week from LightGBM+MetaInfo model