Metro Traffic Time Series Prediction

Chen-Yang Lu

Department of Computer Science National Taiwan University No. 1, Sec. 4, Roosevelt Rd., Taipei 10617, Taiwan (R.O.C.) b02902025@ntu.edu.tw

Wei-Hsi Lin

Department of Computer Science National Taiwan University No. 1, Sec. 4, Roosevelt Rd., Taipei 10617, Taiwan (R.O.C.) b02902035@ntu.edu.tw

Bo-Yao Chen

Department of Computer Science National Taiwan University No. 1, Sec. 4, Roosevelt Rd., Taipei 10617, Taiwan (R.O.C.) b02902071@ntu.edu.tw

Yi-Chen Huang

Department of Computer Science National Taiwan University No. 1, Sec. 4, Roosevelt Rd., Taipei 10617, Taiwan (R.O.C.) b02902053@ntu.edu.tw

**ABSTRACT**

Our goal is to predict the daily passenger flow of each station in Taipei metro system. The metro system is a major public transportation in Taipei, being able to predict the crowds in each station can help people schedule a more relaxed day when they are moving around the city. To achieve our goal, we believe learning from related features and data from the past will help.

# RELATED WORKS

According to "A Brief Survey on Event Prediction Methods in Time Series" [1], Machine Learning Methods and Statistical Learning Methods are two main groups in Time Series Analytics.

## Support Vector Regression

Many researches use SVM to predict future direction of stock price index, so we tried to use SVM for the passenger flow prediction. Support Vector Machine can deal with not only classification problems, but also regression ones (SVR). SVR predict the plane of data distribution, whose label values are continuous corresponding to their feature instances. [2]

## Autoregressive Moving-Average

ARMA is a common way to analyze time series. Autoregressive model uses previous values to predict current value. Moving-average model considers the effect of current and previous white noise errors on current value. ARMA model combines the two models to perform prediction on time series. [1]

# SYSTEM DESIGN

## Algorithm

The algorithm we used to predict the passenger flow will be either SVR and ARMA since the performance of SVR and ARMA can differ on different station. We will first train and predict each station with both algorithms separately, then we choose the better result of the two as the final outcome. Throughout the process, we can observe the more suitable scenario for each algorithm respectively.

## Feature Selection

At first, we considered the influence of work days and holidays, or week days and weekends for most cases, as different locations served different purposes, e.g. business or entertainment. Thus, the difference in work days and holidays can impact the result of some stations greatly.

The second feature we thought of is the influence of weather because we assume people’s choice of public transportation can change depending on the weather.

The last feature we included is the events held nearby the metro stations since large events grab more attentions and people will more than likely attend the event by taking the MRT when there’s a metro station around.

## Data Retrieval

Data from Taipei Metro is available on Taipei Open Data Platform (https://data.taipei). We download the passenger flow data from Jan, 2015 to Apr, 2016, and metro activity calendar from Jul, 2015 to Apr, 2016.

As for weather data, we guess the HTTP 'GET' pattern on 'Weather Underground' (https://wunderground.com/) and get the history weather data in Taipei. It includes temperature, humidity, sea level pressure, etc.

Pandas and numpy in python help us to process data in dataframe structure.

## Data Flow

In SVR, we use weather data, metro stations activity calendar and vacation calendar as our features to train and predict passenger flow (label). ARMA refer the past passenger flow (statistical data) for prediction. Prediction would be send to front-end (web) for data visualization.

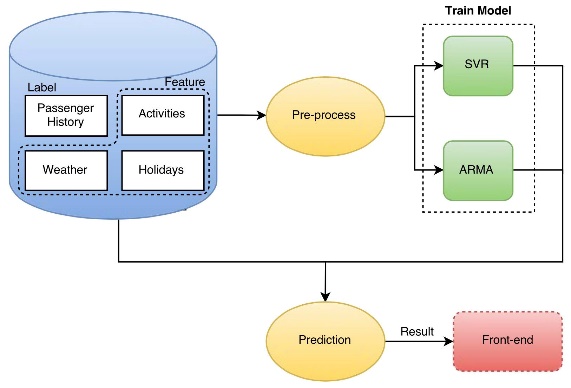


Figure 1: Data Flow Graph

# RESULT

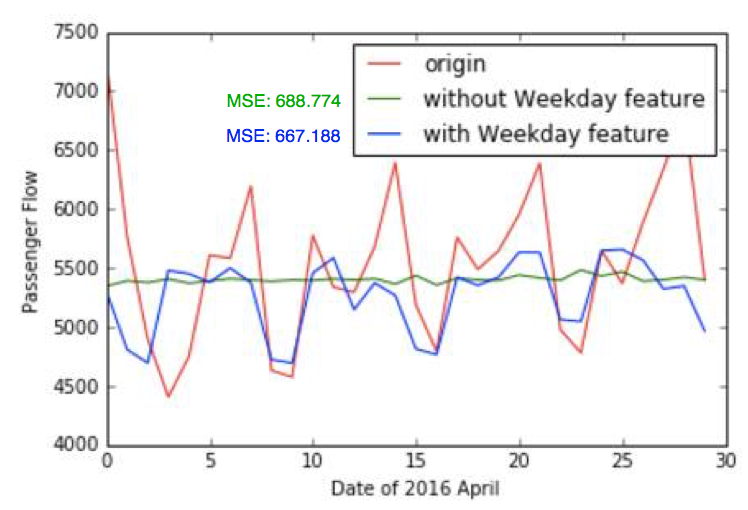
To test our accuracy, we use the data from Jan, 2015 to Mar, 2016 as test data, and predict the passenger flow of Apr, 2016. Then we can compare our prediction with the real data.

We use MSE (Mean squared error of each day's passenger flow) as our estimator:

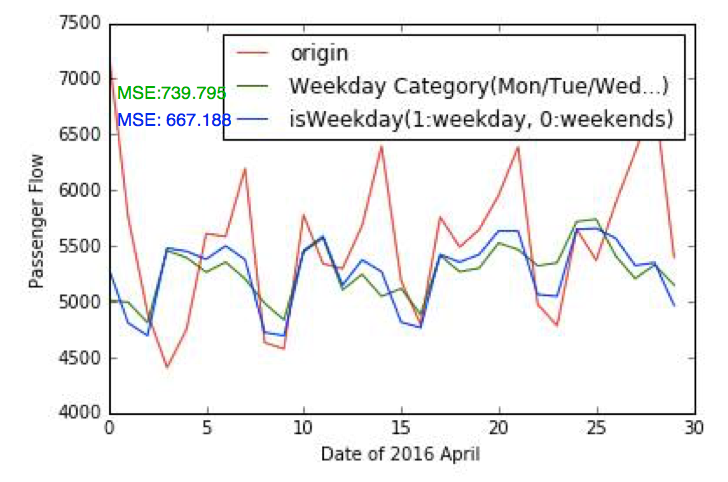
https://scontent-tpe1-1.xx.fbcdn.net/v/t34.0-12/13521153_1068290953261632_1324592191_n.png?oh=d806553c35443b181c0a730ff4e688ac&oe=57721D05

## SVR Feature’s Impact on Result

The following results in this section all used Songshan Airport Station as example. After some observations, we find the weekday or weekend feature to be the most important one.

  
Figure 2: Comparison of prediction with/without weekend feature.

To go further on the weekend feature, we try specifying the weekday, e.g. Monday, Tuesday, of the day as the feature to see if we can get better result. The result, to our surprise, is slightly worse than the weekend feature.

  
Figure 3: Comparison of prediction with the specifying weekday feature and the original weekend feature.

## ARMA Cycle Time’s Impact on Result

For most cases, passenger flow follows a certain cycle pattern every 7 days. For some rare cases, this pattern does not apply, so SVR will perform poorly as SVR mainly depends on the weekend feature we previously mentioned.

In these cases, ARMA becomes a better choice since we can train ARMA with different cycle times. The following result used Zhongshan Station as example.

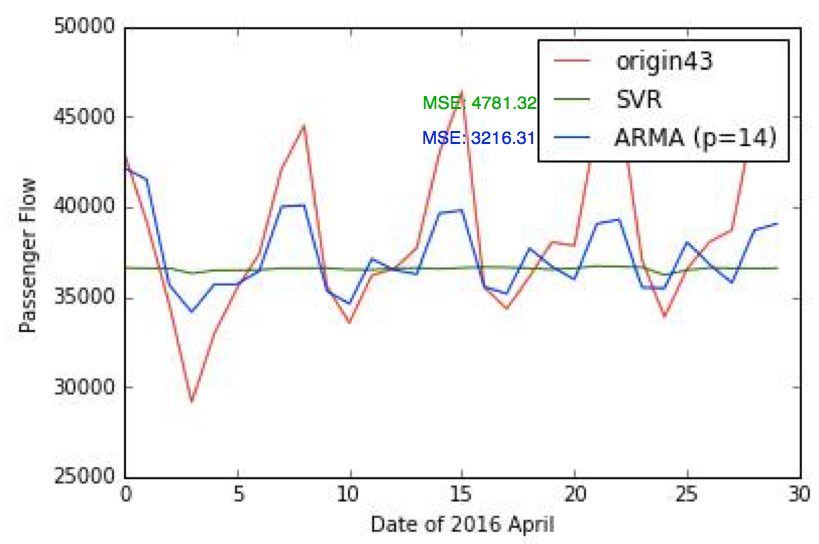


Figure 4: Comparison between SVR and ARMA with cycle time = 14 days

# CONCLUSION

After trying several cases, we come up with the conclusion that cycle time is one of the key element to an accurate prediction. For SVR, the cycle time is fixed to 7 days (the weekend feature). For ARMA, we can train data to find the most ideal cycle time.

We also look into real data to come up with some other conclusions. By observing the data, we find a high point on Dec 31, 2015 and a low point on Aug 8, 2015 for every stations. The reason is pretty simple, they are New Year’s Eve and the day a really strong typhoon struck Taipei, respectively. Interesting enough, they actually belong to our event feature and weather feature. This leads to the conclusion that when considering weather and events, only drastic condition will have noticeable impact on passenger flow. So instead of using them as an everyday feature, we should focus more on them being extreme cases.

# FUTURE WORK

We desperately need more data for features on events prediction, for now our data only involves events right beside the metro stations. This exclude some events that can effectively boost the passenger flow.

Also, for now, the event feature is simply binary, which is without a doubt unrealistic, as there are different scales of events. A way to determine the scale of an event is using search engine to search the event and retrieve the total search results.

# REFERENCE:

[1] Soheila Mehrmolaei. A Brief Survey on Event Prediction Methods in Time Series 10.1007/978-3-319-18476-0\_24, 2015

[2] Zi-heng Huang(R94922044). Support Vector Regression Guide

# CONTRIBUTION

Chen-Yang Lu: ARMA implementation, paper writing

Wei-Hsi Lin: data visualize, part of paper writing

Yi-Chen Huang: data preprocessing, front end

Bo-Yao Chen: data crawling, SVR implementation, data trainin