LIS590DT Final Project:

Data Mining Process for Highway Tollgates Traffic Flow Prediction with Weka 3.8

Wanjie Wang

Jialu Wang

**Background**

Our dataset is from the KDD CUP 2017 Challenge on Highway Tollgates Traffic Flow Prediction to predict travel time and traffic volume. “Highway tollgates are well known bottlenecks in traffic networks, especially during rush hours. Effective preemptive countermeasures such as expediting the toll collection process and streamlining future traffic flow are desired to solve this challenge. For example, if heavy traffic in the next hour is predicted, then traffic regulators could immediately deploy additional toll collectors and/or divert traffic at upstream intersections.”

However, countermeasures will only work when there are reliable predictions for future traffic flow. Traffic flow patterns vary due to different stochastic factors, such as weather conditions, holidays, time of the day, etc. “An unprecedented large amount of traffic data from mobile apps such as Waze (in the US) or Amap (in China) can help us take up that challenge.” If approaches for future traffic flow and ETA prediction can be designed, then the traffic management authorities might be able to capitalize on big data & algorithms for fewer congestions at tollgates.

**Objectives**

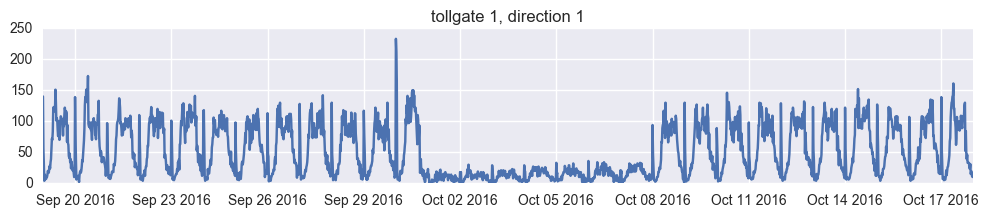
The traffic data set is provided by the KDD Cup but are real data so we can expect a pre-processing process. The goal of the project to explore the data set methodically, distinguish the noise data with data-preprocessing process, and prepare the data set to predict average tollgate traffic volume.

**Data Pre-Processing**

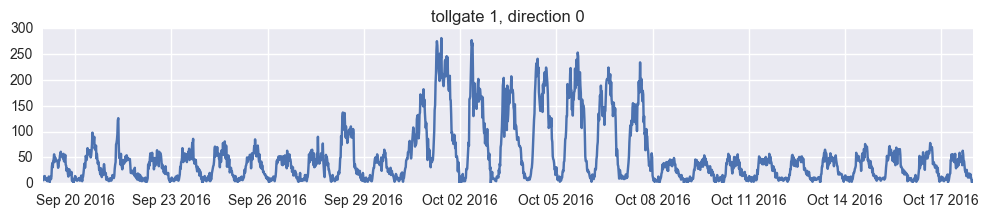
Data pre-processing is an important process in data mining as it determines the quality of input and the results of output. Data pre-processing in this project is conducted with the help of python scripts. The following steps are conducted:

1. Remove the data from October 1st to Oct 7th:

The plot1 and plot2 is the traffic flow at tollgate 1 direction 1 and tollgate 1 direction 0. We find that the traffic volume from October 1st to October 7th has a totally different pattern than that in other time. It is reasonable because in China, the October 1st to October 7th is the National Day of China and people will have a 7 days’ holiday, which will affect the normal daily traffic flow pattern. In addition, the test set only includes normal traffic flow prediction, which means it does not have holiday pattern. So, we remove the data from October 1st to October 7th.



plot 1



plot 2

2. transform the date column into week\_of\_day, hour and minutes:

The data before has a date column includes both date and time value. We transform the date value to week\_of\_day because we think that it is the weekday and weekend will affect traffic flow volume, not the date value. Besides, every week day has a rush hour from 8 to 10 and from 5 to 7. The traffic flow pattern for the rush hour is different from that for other hours. So, we transform the date column into 3 features: week\_of\_day, hour and minutes.

3. use the last 6 time slots volume as features:

Traffic flow is a time series data and the previous time slots will affect the current time traffic flow volume. We read a research paper [1] about using the last time slot traffic volume to predict the current time slots traffic flow. We use the 6 previous time slots traffic flow volume as features.

4. split data into different tollgate\_id and direction

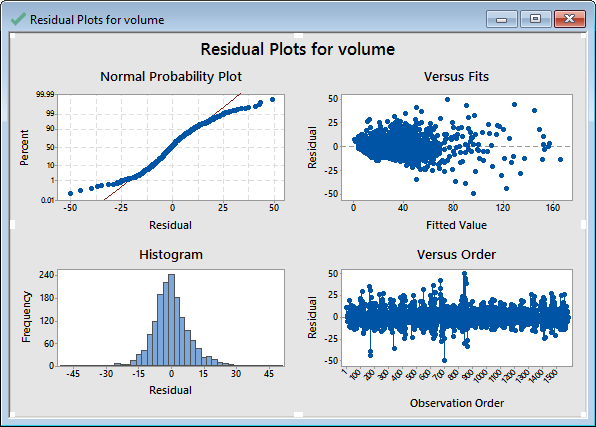
The dataset includes different tollgate and direction together. For each tollgate and direction, the traffic flow has different pattern so we split the data set into 5 parts, each for a (tollgate\_id, direction pair) to improve the accuracy. Tollgate\_id is a binary of 0,1,2 to represent the three tollgates while direction pairs comes in 0 and 1 to represent enter or exit direction of traffic.

**Data Mining Process using Weka**

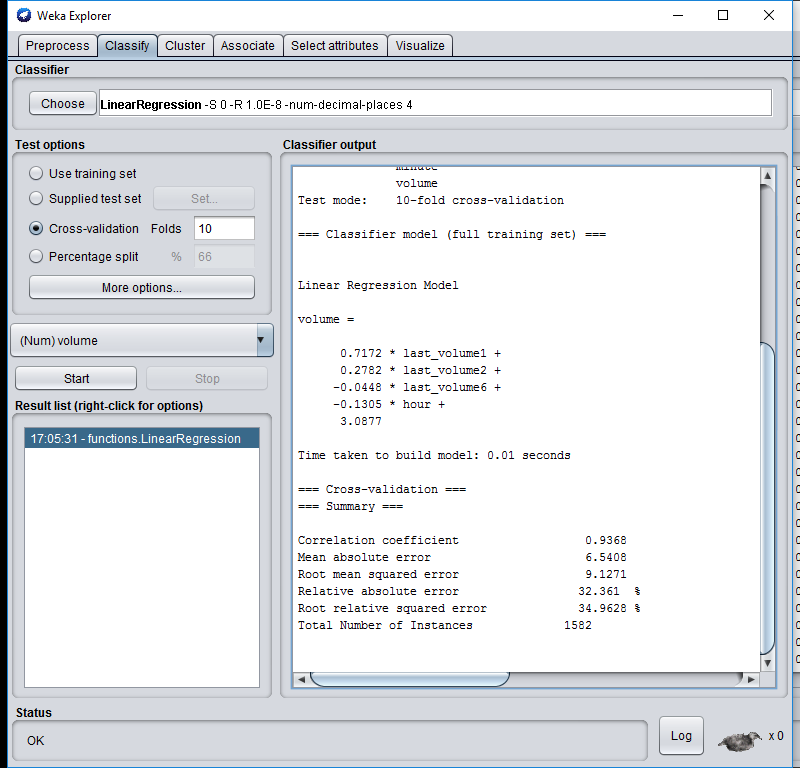
We choose 4 different regression algorithms on each data set and use parameters to compare the results, the following plots are the best results with best parameter on tollgate\_id 1, direction 0.

As our data is numeric instead of text formatted, the selected algorithms are: linear regression, ibk, bagging+M5P, MLP. The reason we choose these 4 algorithms is they performs well in highly related time series data and we take both the linear and non-linear model into consideration. The parameter used is root mean square error (RMSE) to represent for error terms while the accuracy rate is not produced.

1. linear regression:



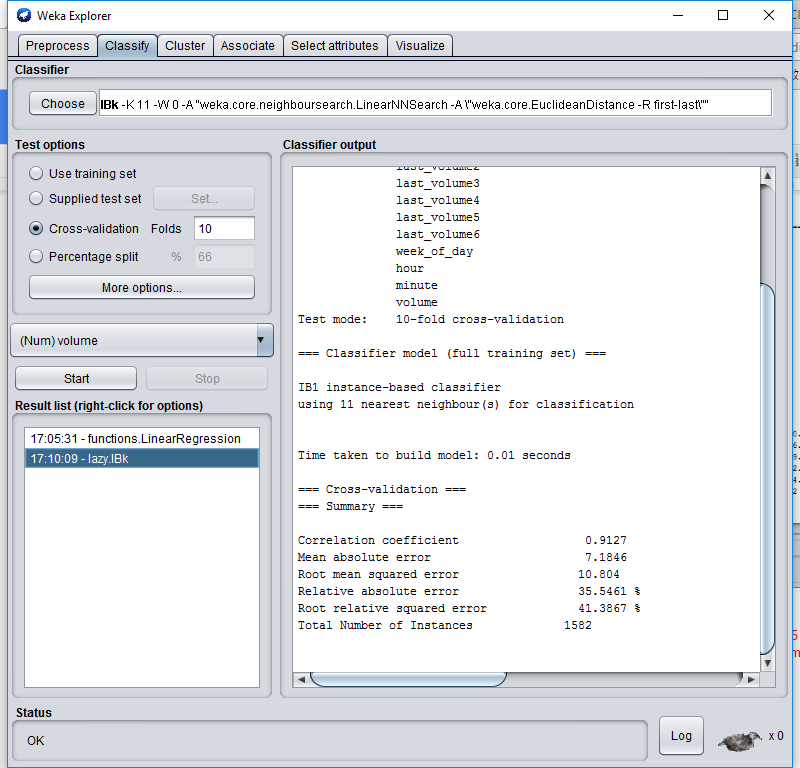
plot 3: linear regression plot

Plot 3 shows the result for linear regression, the residual seems normal but has some degree of heteroscedasticity. The equation shows only last\_volume1, last\_volume2, last\_volume6 and hour 4 features that affect the current traffic flow. The RMSE is 9.1271 for this algorithm. 

plot 4: result for linear regression

1. ibk(k=11)

The plot 5 shows the result for k nearest neighbor regressor(k=11). The RMSE for this algorithm is 10.804 and it is worse than the linear regression. We tried different k values and find the 11 gives us the lowest RMSE.

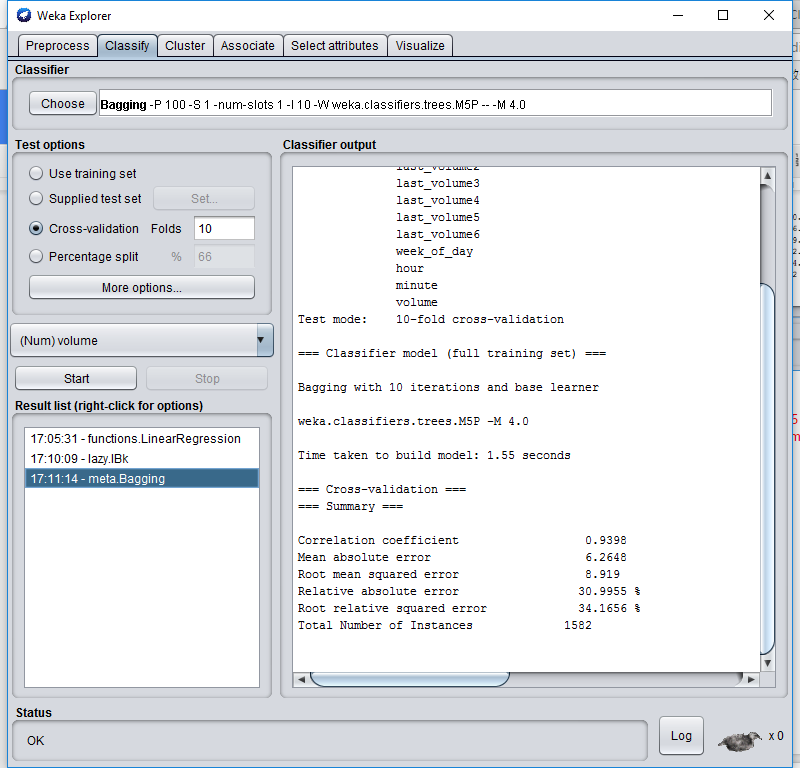


plot 5: result for k-nearest neighbor(k=11)

1. bagging + M5P

M5P is an algorithm similar as NBTree we learned, except that the leaf node of the tree is linear regression.

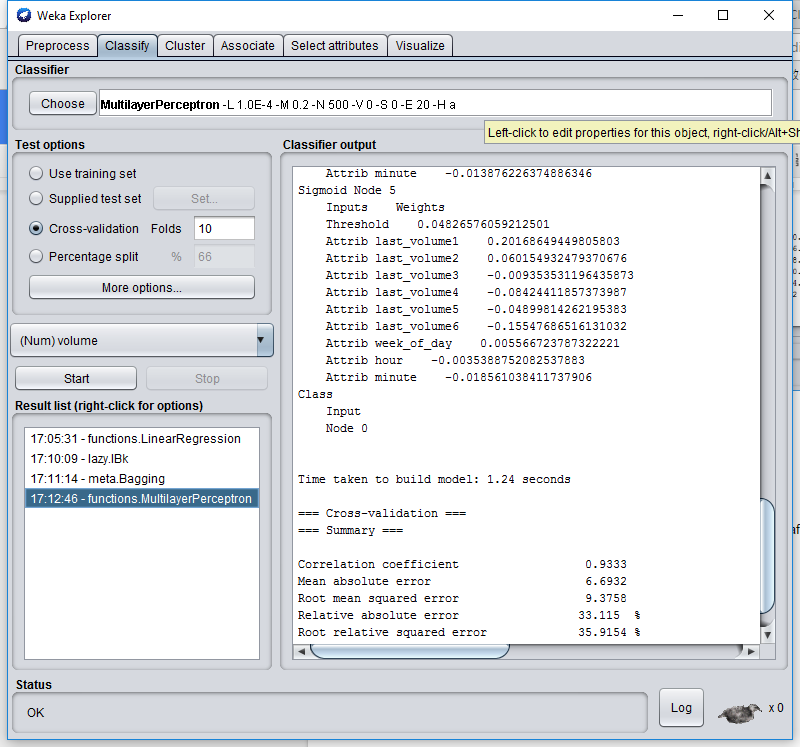
The plot 6 shows the result for bagging+M5P decision tree regressor. The RMSE is good, only 8.919. As we learned in class, the meta learning using decision tree as base estimator gives the best prediction than others. We could see bagging make the model more robust and decrease the RMSE compared with only M5P algorithm without bagging.



plot 6: result for M5Pdecision tree regression with bagging

1. MLP

The plot 7 shows the result for multilayer perceptron. It has 5 hidden layers and each layer using sigmoid as activation function. The RMSE for Multilayer Perceptron algorithm is 9.3758.



plot 7: result for Multilayer Perceptron

**Conclusion and Limitation:**

The M5P decision tree with bagging performs best for predicting current traffic flow. In average, the non-linear model performs better than linear model, which means the data set itself is non-linear and it satisfies our expectation to the data set. There are some research papers using deep learning algorithm like recurrent neural network and Long-Short-Term-Memory RNN to predict the traffic flow and we will try to figure out the model and try to implement to predict the current traffic flow.

**Reference**

Çetiner, B. G., Sari, M., & Borat, O. (2010). A neural network based traffic-flow prediction model. Mathematical and Computational Applications, 15(2), 269-278.

KDD Cup 2017. Retrieved from https://tianchi.aliyun.com/competition/information.htm?spm=5176.100067.5678.2.8CnCPt&raceId=231597&\_is\_login\_redirect=true