Machine Learning based model for short-term traffic prediction

Yuhao Chen
Haonan Peng
Tianmu Wang

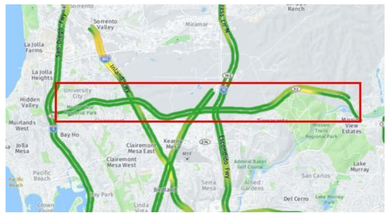
UC San Diego

Abstract

With the expansion of metropolises and the high demand of vehicles, improving the capacity of road networks is no longer simply planning more roads and more lanes. Under that urgent demand, the Intelligent Transportation System (ITS) attracted increasing attention and developed fast in recent years. Among this gigantic system, short term traffic prediction is a vital branch, especially in real-time route guidance. In this project, our goal is to propose a robust and accurate machine-learning based model for short-term traffic prediction.

Introduction

The goal of this project is to do short-term prediction about the traffic flow. The train data is obtained from the PeMS, which shows recorded traffic flow data for the Freeway SR52-E in District 11 (San Diego) for the past few years. In this project, the Gradient Boosting Decision Tree (GBDT), a popular machine learning algorithm in classification and regression field and needs to scan all the data instances to estimate the information gain of all possible split points [1], would be used as Liu [2] did. In this case, the SR52-E connects the University City to the other area, so the prediction of VMT could be used to estimate the rough lane occupancy, which might provide the information to help to schedule the maintenance plan.



Location of the Freeway SR52-E

Procedures

- Preprocessing:

  1) Training dataset obtain:

  The raw training dataset is the traffic flow data of Freeway SR52-E in 2021, which has been download from PeMS, who contains the following features: Month, Date, Hour, Last Hour Vehicle Hours Traveled (VHT), number of Lane Points, percentage of the observed vehicles, and the Last Hour value Vehicle Miles Traveled (VMT). In conclusion, more than 7,000 data points with 7 features are used for building the model, and the prediction should be the VMT for the next hour.

  2) Split of training dataset:

  Since pruning of decision trees requires validation data, the training dataset should be shuffled to remove linearity, and chosen 30 percent of the training dataset randomly by the sklearn function train_test_split

  3) Testing dataset obtain:

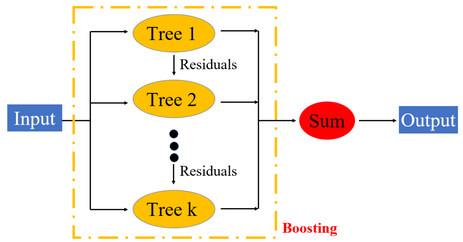
  The raw testing dataset is the traffic flow data of the same freeway in 2022, which would not be exactly the same as the 2021.

- GBDT modeling:

  The GBDT is a Ensemble Model, who is to combine several models (commonly known as weak estimator) according to a certain strategy to jointly complete a task, then a specific combination strategy would help the ensemble model reduce the bias or variance of predictions. The GBDT is using the boosting strategy.

  1) Boosting Strategy:

  Boosting models combining the basic models by concatenation. The idea of this type of model is that one basic model can make imperfect predictions, so more models could be applied to polish the imperfect parts. And the actual structure would be like:



Boosting Model Schematic

  2) Classification and Regression Tree (CART)

  GBDT is a boosting combination of k's CART, so CART is the basic model. In such regression problem, the binary tree should be used to complete linear regression tasks. To construct the tree, the algorithm will first generate some split filters according to the Gini impurity, then the variables that satisfies the filters would be dropped to the left branch while the rest in the right branch. The split would keep processing until the stop criteria reached (like the maximum tree depth). For the regression tree, the Gini impurity would be replaced by the variance reduction, which is used to increase the precision of the estimations by:

$$VE = \frac{1}{N}\sum_{n=1}^{N} Var_n - \left(\frac{length_{left}}{length} \times Var_{n_{left}} + \frac{length_{right}}{length} \times Var_{n_{right}}\right)$$

  Then after building the model, the validation dataset would be used to test the model and the tree would be pruned according to reduce error pruning (REP), so the model would eliminate the overfitting branches.

  3) Gradient and Loss function:

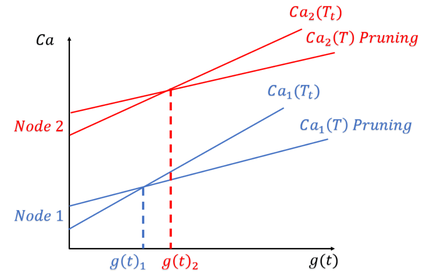
  The gradient is used to accelerate the process of minimizing the loss function, which is the criterion how the algorithm judges the quality of training. In this case, the Huber loss, a loss function used in robust regression and it is less sensitive to outliers in data than the squared error loss, and the relative square loss gradient applied to construct the boosting model. And in this case the sigma is equal to 1.

$$L_\delta(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2, & if\ |y - f(x)| \le \delta \\ \delta \cdot \left(|y - f(x)| - \frac{1}{2}\delta\right), & otherwise \end{cases}$$

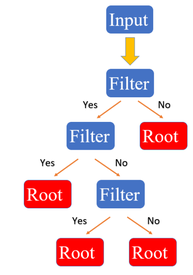
$$Gradient_\delta(y, f(x)) = \begin{cases} y - f(x), & if\ |y - f(x)| \le \delta \\ \delta \cdot sign(y - f(x)), & otherwise \end{cases}$$

- Prediction:

  After training and pruning the model, the prediction could be done by input the variables. Here, in order to verify the correction of the model, some testing dataset should be used then output the prediction, which could be compared to the actual values and see how accurate the result is



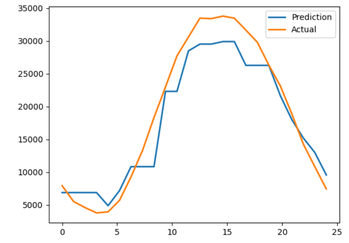
Error Difference Before and After Pruning



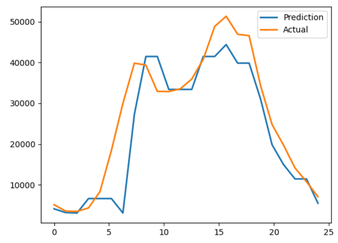
CART Schematic

Results

The results from two random day in 2022, the tendency of the prediction would be similar to the actual ones although the prediction might not be smooth and accurate enough.



GBDT Prediction



GBDT Prediction

Future Work

In conclusion, the GBDT algorithm could predict the traffic flow basing the given data, which might help to adjust the strategy traffic control. Nevertheless, the variables using in this project is not quite enough, many parameters like the weathers, vacations, price of gasoline, are no included, so the prediction may not accurate enough. Therefore, the improvement of the project should be concluding more potential parameters in the model, so as to improve the accuracy. Also, more accurate model might provide a more robust long-term prediction

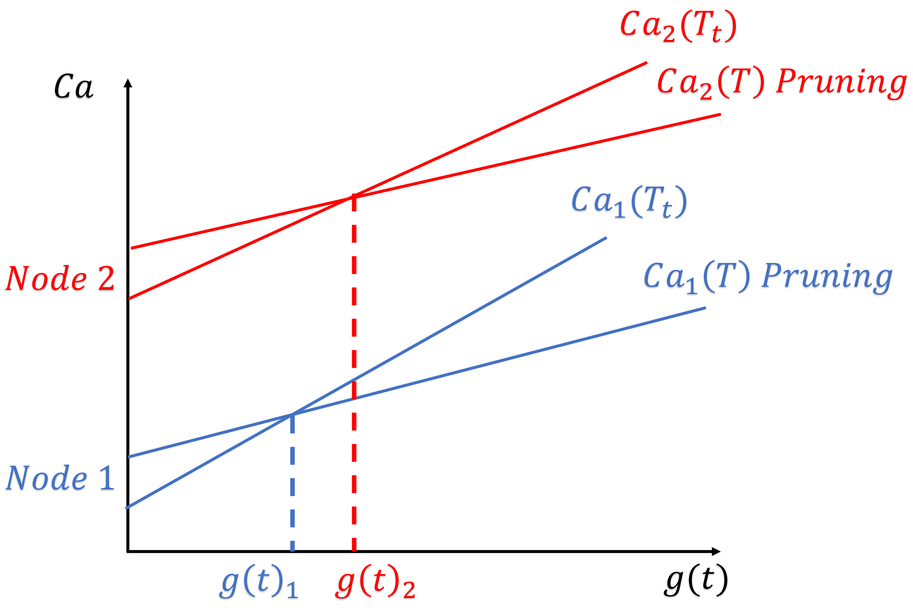
References

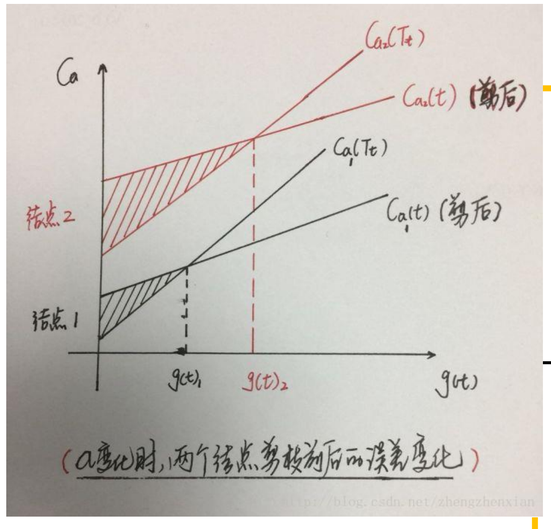
[1] Ke, Guolin, et al. "Lightgbm: A highly efficient gradient boosting decision tree." Advances in neural information processing systems 30 (2017).

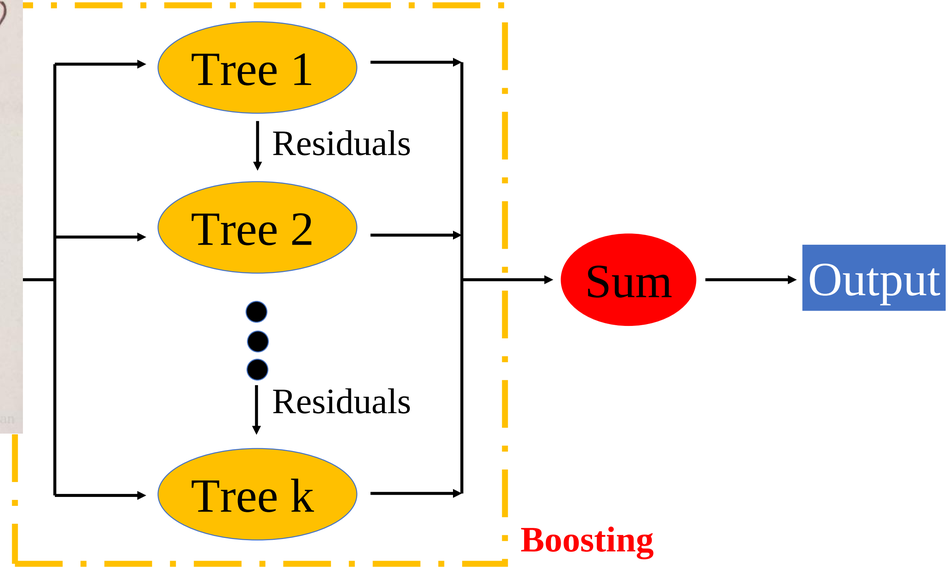
[2] Liu, Yingxin, et al. "Traffic Flow Forecasting Analysis based on Two Methods." Journal of Physics: Conference Series. Vol. 1861. No. 1. IOP Publishing, 2021.

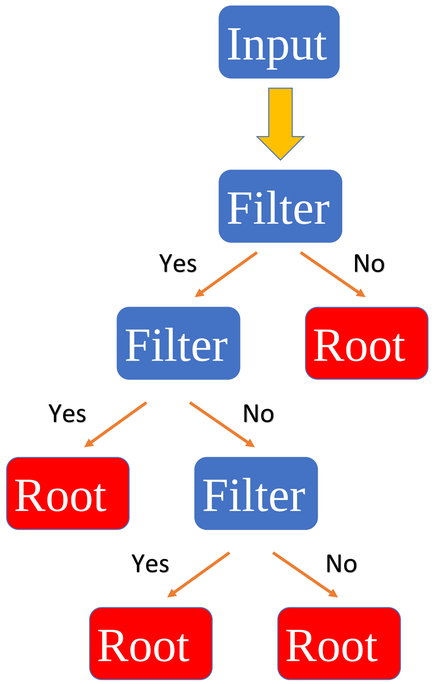
[3] Natekin, Alexey, and Alois Knoll. "Gradient boosting machines, a tutorial." Frontiers in neurorobotics 7 (2013): 21.

[4] Boukerche, Azzedine, and Jiahao Wang. "Machine Learning-based traffic prediction models for Intelligent Transportation Systems." Computer Networks 181 (2020): 107530.


Ca
$Ca_2(T_t)$
$Ca_2(T)$ Pruning
$Ca_1(T_t)$
$Ca_1(T)$ Pruning
Node 2
Node 1
$g(t)_1$
$g(t)_2$
$g(t)$


Ca
$Ca_2(T_t)$
$Ca_2(t)$ (剪后)
$Ca_1(T_t)$
$Ca_1(t)$ (剪后)
结点2
结点1
$g(t)_1$
$g(t)_2$
$g(t)$
(a变化时，两个结点剪枝前后误差变化)


Tree 1
Residuals
Tree 2
Residuals
Tree k
Sum
Output
Boosting


Input
Filter
Yes
No
Filter
Root
Yes
No
Root
Filter
Yes
No
Root
Root