Machine-learning Based Model for Short-term Traffic Prediction

Yuhao Chen
Department of MAE
yuc043@ucsd.edu

Tianmu Wang
Department of MAE
tiw028@ucsd.edu

Haonan Peng
Department of ECE
hap045@ucsd.edu

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. Besides, the increasing number of roads is unavoidably raising the complexity of road networks which brings difficulty in monitoring and maintaining the traffic environment. 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.

1 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 GBDT model would be implemented basing on the **Hour** (H), the **Vehicle Hours Traveled** (VHT), the **number of Lane Points** (N), and the **percentage of the observed vehicles** (O), then the **prediction of Vehicle Miles Traveled** (VMT) for each hour through a day could be obtained. 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.

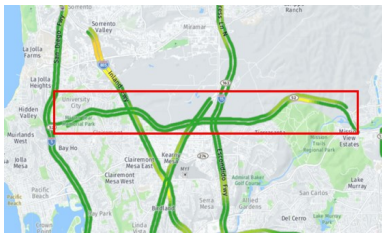


Figure 1: Location of the example freeway

2 Methodology

The **Gradient Boosting Decision Tree** (GBDT) method would be applied to construct the prediction model.

GBDT is also known as **MART** (Multiple Additive Regression Tree), it is for classification and regression problems. The algorithm could train the newly added Weak Classifier according to the negative gradient information from current loss function, then combine the well-trained weak

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Table 1: Statistics of traffic data sample

Hour	Vehicle Miles Traveled	Vehicle Hours Traveled	Lane Points	%Observed
4/1/2022 0:00	5,260.90	79.10	480	55.6
4/1/2022 1:00	3,916.50	59.30	480	55
4/1/2022 2:00	3,683.00	55.60	480	55
4/1/2022 3:00	4,486.30	67.90	480	55
4/1/2022 4:00	7,591.00	113.50	480	55

classifier with the existence model by accumulation. Then, since it is a decision-tree-based algorithm, we would discuss the decision tree first.

Decision tree algorithm is a typical classification method to approximate the value of a discrete function [1][2]. First, it would process the data using an inductive algorithm to generate the readable rules and the decision trees. Then, it would use the decisions to analyze the new data, so as to achieve the goal of classification or regression. Thus, essentially, the decision tree is the process of classifying data through a series of rules.

The classic algorithms of the decision tree are ID3, C4.5, and CART.

In this project, the **CART** (classification and regression tree) is used to serve as the decision tree in the GDBT algorithm, because the CART is available for the regression problems, which helps to provide a prediction of traffic flow. It is a non-parametric decision tree learning technique that produces either classification or regression trees, depending on whether the dependent variable is categorical or numeric.

The representation of the CART model is a binary tree, and each root node represents a single input variable and a split point on that variable. As for the leaf nodes, they stand for the output variables, and are used to make a prediction for the new data.

3 Data Description

The data we used in the study is downloaded from the Caltrans Performance Measurement System (PeMS) (`https://pems.dot.ca.gov/`), which is an open-access, real-time database collecting the traffic data of the freeway system across all major metropolitan areas of the State of California from over 39,000 individual detectors. We collected the traffic data of the State Route 52 in San Diego County, California from January 1 to April 30, 2022. The selected segment is a major east-west route with entire length 14.8 miles monitored by 44 detectors. In our study, the data we collected is grouped by: hour (represented in 24-hour clock), vehicle hours traveled, vehicle miles traveled (the product of vehicle hours traveled and entire length of route), lane point, and the percentage of observed vehicle. The detailed statistic sample is shown in Table 1. The traffic data of January, February, and April is selected to train our model, while the validation data is the data collected on March. It should be mentioned we selected the data on March 1st to test our model at the current stage.

4 Experiment

The general training process is as follows. First, it would select the features from the given training dataset, and get the best split to differentiate the observations based on the dependent variables. For example, in this project, the prediction variable is VMT, and thus the features would be H, VHT, N, and O. Then, under one feature, the binary decision tree would be applied to divide the space, which is a numerical process, and the cost function and different split nodes help to select the optimal split. For the classification, the Gini Index function would be applied as an indication, which describes how mixed training data assigned to each node is [2]. As for the regression predictive modeling problem, the sum squared error would become the criterion that needs to be minimized so as to obtain the optimal split.

Thus, when the instances in the feature belong to the same class, then it would be set as a single node. Otherwise, the split would begin, the instance would compute the Gini impurity once it is less than the threshold, the recursion would then stop and return the children node. Then, if the Gini impurity

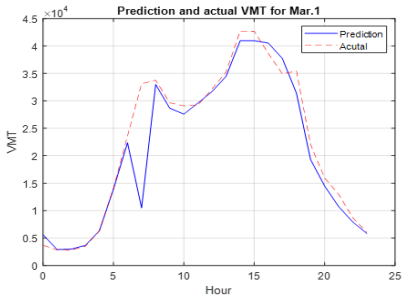
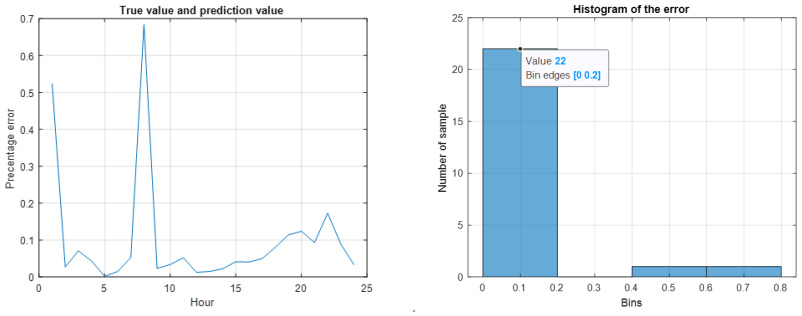


Figure 2: prediction vs. ground truth



(a) (b)

Figure 3: Prediction error on: a) percentage, b) histogram

is larger than the threshold, it would compute all features for this node, and obtain the minimum one as the optimal division, and it would go left or right depending on if it satisfies the conditions. Through this way, the CART classification tree would be well-trained, and the test data would then be predicted according to this decision tree. The test sample would follow the decision tree and finally stay in some children nodes, and the corresponding desired prediction value, VMT in this project, would be returned to the test samples as prediction.

So far, we built up a training CART model based on the dataset from the PeMS, which includes the January, February, and April of 2022. Then, the test data is from recording of March 1st, 2021. The results are shown in Figure 2.The general tendency of the prediction is quite similar as the actual data, and the relative error and the histogram of error are shown in Figure 3. Then, if we consider the prediction with error less than 20% as the proper prediction, then the accuracy of this prediction is 91.67%, which seems good but there still exists some large error. Thus, this prediction model can still be improved to be more accurate.

However, if we try to predict for a longer period, the performance of the model would be much worse. Figure 4 demonstrates the case for predicting the VMT for the whole July of 2021, and the corresponding results are as follows. Among 744 testing samples, only 442 had the acceptable error, so the accuracy would be only 59.41%, which seems unacceptable. In other words, it shows that our model is not suitable for prediction on long term traffic flow.

5 Conclusion

At the current stage, the conducted experiment proves that CART decision tree can handle the short-term prediction on traffic flow well while giving us space for improvement. For further study, our first goal is to complete the gradient boosting algorithm of our ensemble architecture to enhance the stability of model and its accuracy on prediction. Meanwhile, we will also continuously explore the potential features of the data to come up with a best sets of features for training. At last, we plan to introduce several neural network models that are also popular on short-term prediction, such as

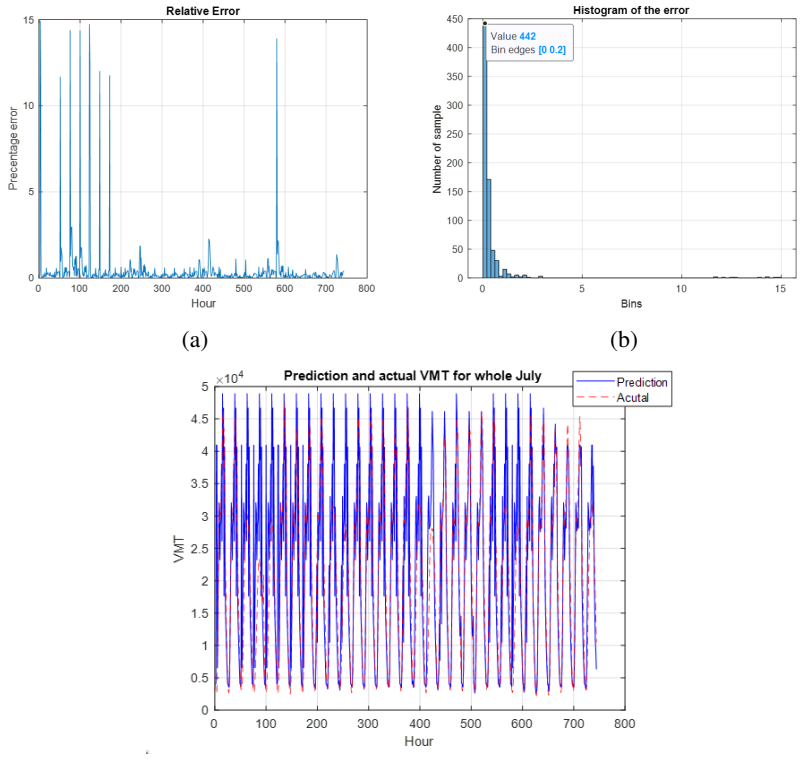


Figure 4: Long term prediction: (a) Relative error, (b) Histogram of error, (c) VMT for whole July

LSTM and GRU [4], to comprehensively evaluate the performance of each model to point out the advantage and drawback of our design.

References

[1] Loh, Wei-Yin. "Classification and regression trees." Wiley interdisciplinary reviews: data mining and knowledge discovery 1.1 (2011): 14-23.

[2] De'ath, Glenn, and Katharina E. Fabricius. "Classification and regression trees: a powerful yet simple technique for ecological data analysis." Ecology 81.11 (2000): 3178-3192. Dataset:

[3] Senyan Yang, Jianping Wu, Yiman Du, Yingqi He, Xu Chen, "Ensemble Learning for Short-Term Traffic Prediction Based on Gradient Boosting Machine", Journal of Sensors, vol. 2017, Article ID 7074143, 15 pages, 2017. https://doi.org/10.1155/2017/7074143

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