Research on traffic travel time prediction based on machine learning

PeiTing Zhang

School of Economics and Management, Beijing JiaoTong University, Beijing, China
20120594@bjtu.edu.cn

ABSTRACT. With the rapid development of economy, the pressure of traffic operation is increasing, and the improvement of intelligent transportation system is more urgent. In this paper, we focus our research on the topic of travel time prediction based on machine learning, and use the international data mining field event dataset as the research sample. First, the data are initially analyzed to extract the feature variables; then the prediction of travel time is further performed by using linear regression, ridge regression and random forest regression to compare the evaluation results. Based on the data test results, it is found that linear regression can achieve better prediction effect than ridge regression and random forest in the traffic data environment of highway. It can be seen that in the intelligent transportation system, although the complex model can improve the prediction accuracy, it does not necessarily achieve better prediction effect, which provides a reference basis for the selection of driving time prediction method in the transportation system.

Keywords: Travel time, Regression algorithm, Random forest

1 INTRODUCTION

Along with the rapid development of science and technology such as "artificial intelligence" and "Internet of Things" and the continuous improvement of people's living standards, people's demand for transportation services has gradually changed. Today's logistics and transportation services are not only satisfied with the basic requirements such as on-time and safe delivery, but also begin to focus on how to establish a more perfect "intelligent transportation system", so that the development of logistics can respond more effectively to the changes in people's life needs.

This study will focus on the prediction of travel time in smart logistics, first of all, we need to clarify what is smart logistics? In this paper intelligent logistics refers to the intelligent movement process of goods from the supply side to the demand side, including six basic activities such as intelligent storage, intelligent transportation, intelligent distribution, intelligent packaging, intelligent loading and unloading, and intelligent collection, processing and handling of information within. The prediction of travel time in this study can effectively explore which factors affect travel time, and also use the technology of machine learning to build a model to provide effective support for the construction of intelligent transportation system.

V. Escudero et al. (Eds.): BADS 2022, AHCS 11, pp. 893–898, 2023.
https://doi.org/10.2991/978-94-6463-102-9_92

In the existing studies on traffic congestion, the first concern is which factors may affect travel time, and most of the studies focus on a single dimension, such as only the length and width of road sections, and few studies focus on the effects of multidimensional factors such as time, road, and weather on travel time. On the other hand, when evaluating model effectiveness, most studies choose to use multiple datasets to validate a certain algorithm, and few studies focus on the performance of different types of machine learning algorithms in the same dataset, and if they do compare, they do so only by internal variations such as tuning parameters.

2 Data source and description

In order to better examine the application of machine learning algorithms in travel time prediction, the competition data of KDD CUP 2017 was selected as the experimental data sample for this study. The "KDD CUP" competition is an international competition sponsored by the data mining section of the American Computer Society ACM, which invites experts in various fields such as big data and machine learning from around the world by providing data and competition topics.

The data used this time describes the flow of vehicles at highway toll points, mainly including (1) basic information on paths; (2) information on vehicle passage times; (3) data information on three aspects of weather information and road network topology maps of highways and stations. Since the weather factor is a key factor affecting vehicle travel time in different studies, and the time span of the adopted data set is small, the weather factor will not be analyzed and studied in depth. The data on the path and vehicle travel conditions will be presented below.

The route data is the basic data of this study, which contains two parts: route situation and route relationship. There are seven types of features described in the road situation data, including route length, number of lanes, number of intersections, etc.; the road relationship data provides the connection order between intersections and toll stations.

Table 1. Path case variable[Owner-draw]

Variable Name	Variable Description
link_id	Route number of the connection from the junction to the toll station
length	Length of each line passed from the road junction to the toll station (m)
width	Width of the connecting line from the junction to the toll station (m)
lanes	Number of lanes
in_top	Number of intersections connected at the time of entering the toll station
out_top	Number of intersections connected when exiting the toll station
land_width	Lane width
Intersection_id	Intersection number
Tollgate_id	Toll Station Number
Link_seq	Connection order

In the path case data, it can be seen that the data of this study, the longest value interval of the intersection to the toll station line is [6,293], the path width value interval is [3,12], the path lane number interval is [1,4], the overall section in the lane width are 3 meters, the change in the number of lanes in different sections directly affect the path width.

The specific link relationship of the paths is shown in Table 4-3, where intersection indicates the number of each intersection, tollgate indicates the number of each toll station, and link_seq indicates the link order of the paths. From the information in the table, we can see that the data set contains a total of six routes, respectively, A2, A3, B1, B3, C1, C3, each route contains more than one road, and in the order of the links in the table, does not change with time, for example, the A2 route needs to pass through a total of six roads in turn.

The competition provides two sets of vehicle travel time information data, the training set and the test set, where the training data contains the vehicle travel from July 19, 2016 to October 17, 2016, and some of the training data are presented in Table 4-4 due to space limitation. It contains data such as vehicle vehicle number, starting_time vehicle entry line timestamp, and travel_time vehicle travel time (in seconds) experienced from the intersection to the toll booth; the test data contains the vehicle operation from October 18, 2016 to October 24, 2016, etc.

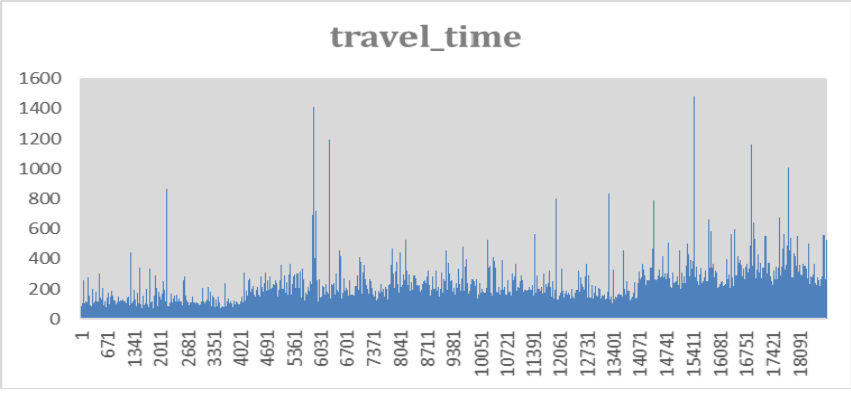

Fig. 1. Travel Time Scatter Chart[Owner-draw]

3 Results

This paper chooses to adopt Python as the main programming language, the advantage of which is the flexibility to call the toolkit. This study uses the sklearn toolkit, and also combines Numpy, pandas, etc. to realize data visualization.

Sklearn modeling process is basically divided into three steps, firstly, the parameters to be applied are instantiated to build the evaluation model; then the model is trained through the interface, and finally the data information is extracted through the interface. The training data set is divided into a training set and a test set before the

library operation, and this study uniformly uses 25% of the data as the test set and 75% as the training set.

3.1 Linear regression

In the sklearn module library, the classifier linear_LinearRegression of linear regression, which mainly contains the following parameters: fit_intercept is used to calculate the intercept of the model, and the cutoff is not calculated by default; copy_X is used to ensure that the original feature matrix is not covered.

The model is also evaluated using the cross_val_score package, and the data are cross-validated with 3-fold using cross-validation to derive the model coefficients and error ratios. The table presents the coefficient values of the eight feature variables, and it can be seen that route selection has the greatest degree of influence on travel time prediction; however, the coefficient of the feature variable of whether or not it is a weekend is only 0.47, and the reason for this phenomenon is likely due to the fact that this study uses a highway dataset. Short-term weekend holidays do not induce a large direct intercity movement of people, so weekends have a small impact on traffic status; among the road characteristic factors, the greater the number of connections in the route, the shorter the travel time

Table 2. LR model coefficients[Owner-draw]

Route_N	length	width	in_count	out_count	link_count	Week	weekday
6.66	0.13	-0.98	2.66	2.09	2.84	0.56	-0.45

3.2 Ridge regression

Based on the basic theory of ridge regression, it is known that ridge regression can effectively solve the problem of multicollinearity in feature values compared to linear regression. The coefficients are shown in the following table. By comparing the presentation of LR and RR regression coefficients, it can be found that there is no multicollinearity in the highway vehicle travel time data selected for this study. In the ridge regression, the coefficients do not change significantly, and only the coefficient of the variable whether it is a weekend (weekday) fluctuates significantly. The main reason for the analysis is that this variable is presented in the form of Boolean values, which leads to the influence of regularization on the coefficient values

Table 3. RR model coefficients[Owner-draw]

Route_N	length	width	in_count	out_count	link_count	Week	weekday
6.91	0.13	-0.93	2.84	2.20	2.80	0.46	-0.08

3.3 Random Forest Regression

The Random Forest Regressor model in Sklearn is called for regression prediction of random forest. Unlike linear regression and ridge regression, random forest does not have fixed model coefficients because multiple models are built at the same time. The parameters in RF are n_estimators, min_weight_fraction_leaf, etc. The three parameters focused on in this study are: the number of base models n_estimators; the maximum tree depth max_depth; and the optimal feature proportion max_features.

In this study, the regression for random forest is mainly done in the following aspects

(1) Random extraction of all samples using the Train_test_split tool, of which 75% is used as the training data set.

(2) Thresholding of three key parameters, n_estimators base model number set to 5,10,20,50,100,200; max_depth maximum depth set to 3,5,7 three depth levels; max_features optimal feature ratio set to 0.6, 0.7, 0.8 and 1.

(3) 3 times cross-validation of the data to ensure the randomness of data sampling

(4) The optimal random forest model is calculated using GridSearchCV, and one of the decision trees is visualized

The experimental results show that the optimal random forest model structure is a maximum tree depth of 5 with a maximum feature selection of 0.6, and the number of base models is 50.

4 Conclusions

The focus of this paper is to study regression algorithm and random forest in vehicle travel time prediction problem. Comparing the two algorithms, the regression algorithm has the advantages of strong model interpretability and low requirements for computer operation ability; the random forest can more effectively take advantage of the integration of the model and exclude the influence of missing values in the data on the model prediction, but the model is less interpretable because the simulation process of the decision tree is completely completed by computer learning.

In the research process of exploring vehicle travel time prediction, eight feature values were selected for travel time prediction using KDD CUP 2017 race road data. It was found that, among them, road selection, the number of connecting sections and the number of toll station exits have a large impact on vehicle travel time, and relatively speaking, the time factor has a small impact on the prediction. After analysis, the results of this study are consistent with the actual situation. Firstly, the data selected are from highways, and short-term weekend holidays do not affect significant changes in inter-city traffic conditions, so the time factor has a small impact on the prediction of highway travel time; secondly, the road factor, the choice of different traffic roads directly affects the length of the trip, which in turn affects the time; finally, the toll booths on highways are an important part of traffic congestion, and the number of toll booths.

The significance of this study is based on the current development needs in the field of intelligent logistics, using machine learning for the prediction of vehicle travel

time to provide theoretical support for further effective guidance of traffic and improvement of logistics service quality.

References

1. Mori, U., Mendiburu, A., Alvarez, M., et al. (2015). A review of travel time estimation and forecasting for Advanced Traveller Information Systems [J]. Transportmetrica A: Transport Science, 11(2).
2. Qiao, Haghani A., Shao C F, et al. (2016). Freeway path travel time prediction based on heterogeneous traffic data through nonparametric model [J]. Taylor & Francis, 20(5).
3. Xu,T., Li,X., Claramunt, C. (2018). Trip-oriented travel time prediction (TOTTP) with historical vehicle trajectories [J]. Frontiers of Earth Science, 12(2).
4. Chen, C-H. (2018). An Arrival Time Prediction Method for Bus System[J]. IEEE Internet of Things Journal, 5(5): 4231-4232.
5. Hoerl, A.E., Kennard, R.W. (2012). Ridge Regression: Biased Estimation for Nonorthogonal Problems [J]. Taylor & Francis Group, 12(1).
6. Zito, R. (2005). A review of travel-time prediction in transport and logistics[J]. Proceedings of the eastern asia society for transportation studies,21(5).
7. Nanthawichit,C., Nakatsuji, T., Suzuki, H. (2003). Application of Probe-Vehicle Data for Real-Time Traffic-State Estimation and Short-Term Travel-Time Prediction on a Freeway [J]. Transportation Research Record, 1855(1).
8. Breiman, L. (1996). Bagging Predictors [J]. Machine Learning, 24(2).
9. Breiman, L. (2001). Random Forests [J]. Machine Learning, 45(1).
10. Zhang, X., Rice, J. A.(2003). Short-term travel time prediction [J]. Transportation Research Part C, 11(3).