Bus arrival time prediction based on Random Forest

Li Jian[1, a]

[1]School of software engineering, Beijing University of Technology, Beijing 100124, China

[a]lj201425005@emails.bjut.edu.cn

Keywords: intelligent transportation system, bus arrival time prediction, random forest.

Abstract: In order to balance the traffic supply to meet the citizens demand for public transportation, reduce the pressure of urban traffic, enhance the competitiveness of public travel and improve the intelligent transportation system service, this paper proposes a bus arrival time prediction algorithm based on Random Forest. In this paper, traveling data of the 607 bus in Beijing is analyzed, the data are pretreated by using Space Rectangular Coordinate System instead of the traditional GPS Geodetic Coordinate System. The traffic junction number, travel distance, date type, time period, precipitation, visibility Six kinds of influencing factors were utilized to model the bus arrival time prediction model using Random Forest. The experimental results demonstrate that the mean absolute percentage error of the algorithm is 20.43% when under the condition of setting 800 decision trees.

1. Introduction

The problem of traffic congestion seriously affects the daily life of the citizens and becomes the bottleneck restricting the development of the city [1, 2]. The accurate bus arrival time prediction algorithm can be applied to the intelligent transportation system to help resolve the problem of the traffic congestion problems. Because of Complex traffic environment, the arrival time of the bus cannot be predicted by a simple model. Therefore, it is a great challenge to research a bus arrival time prediction system which can adapt to complex traffic environment.

In response to this problem, a number of researches have been proposed. According to the literature [3, 4], bus route will be divided into several sections, using historical data to calculate the average travel time of each sub-route and accumulated to get the final forecast time. In [5], the average dwell time is introduced, and the predicted time is composed of the average travel time and the average dwell time. The above two algorithms only use the historical travel data to calculate the average travel time as the predict value. However, the model applies only to the road with stable traffic conditions and cannot cope with the emergency. In [6, 8] dwell time, running speed, route length and passenger volume as influencing factors using Neural network algorithm to construct the prediction model. Although the model takes into account a variety of factors, but the neural network model is easy to over fit, in order to solve this problem needs to pay a high model training costs. In the literature [9], the historical travel data of the bus are divided into four groups according to period and weather (peak period and sunny day, peak period and rainy day, off-peak period and sunny day, off-peak period and rainy day), using support vector regression machine to do prediction. The model can be used to map the input parameters to high-dimensional space and solve the nonlinear problem by selecting the appropriate kernel function. However, the algorithm is simply to separate the weather and time based on the experience. Lack of analysis and evaluation of historical data, cannot be applied to the complex traffic conditions.

Aiming at the above problems, this paper proposes a bus arrival time prediction algorithm based on Random Forest. The algorithm uses the Space Rectangular Coordinate system instead of the Geodetic Coordinate System to pretreat the geographical location of bus, taking into account the traffic junction number, travel distance, date type, time period, precipitation, visibility 6 influencing factors, using the random forest algorithm to model the data. The experimental results prove the accuracy of the algorithm.

2. Technical foundation

2.1 Data preprocessing

Bus location data are collected by the GPS device, which is installed in the bus, collected in real time. But the location of the bus be showing offset from main road when using the data to mark the bus in map directly because of error of the GPS device. In order to eliminate this error, the need for geographical location data to the main line projection calculation, so that the location data of the bus are close to the actual driving route.

In addition, the original GPS data is a Geodetic Coordinate System. The coordinate system is based on the centroid of the earth. It is an ellipsoidal coordinate system. Although it can be approximated in a very small range, it is still cause error when calculating in this coordinate system directly. The original geodetic coordinate system needs to be transformed into Space Rectangular Coordinate System in data preprocessing. The coordinate transformation formula (X, Y, Z) = h (B, L, H) is as follows:

$$X = (N + H) * \cos B * \cos L$$
$$Y = (N + H) * \cos B * \sin L$$
$$Z = [N * (1 - e^2) + H] * \sin B$$

N: ellipsoid curvature radius
e: First eccentricity of ellipsoid
a: major axis of ellipsoid
b: minor axis of ellipsoid
B: geodetic longitude
L: geodetic latitude
H: geodetic height

2.2 Decision Tree

Decision tree is a tree structure, composed of nodes and branches, and is based learning unit of Random Forest. The nodes include 2 types: internal nodes and leaf nodes. Depending on the test results, the samples are assigned to their child nodes, and each sub-node corresponds to a value or range of the attribute, starting from the root node and testing the attribute of the sample. The samples are tested recursively until the leaves are reached, and the samples are allocated to the leaves [10].

The decision tree used in this paper is the Classification And Regression Tree (CART tree) that can be used for classification and regression problems. The regression tree algorithm for regression problems is described as follows:

The dataset is defined as follows:

$D = \{(x_1, y_1); \cdots; (x_i, y_i); \cdots; (x_m, y_m)\}$; m: the number of dataset

(x_i, y_i): the i-th sample of sample

x_i: the feature of the i-th sample

y_i: the output of the i-th sample

The sample feature set is defined as follows: It contains discrete feature (such as time periods) and continuous feature (such as precipitation, travel distance).

$$A = \{a_1, a_2, \cdots, a_n\}$$

n: the number of features

Each sample $(x_i, y_i), 1 \leq i \leq m$ is defined as follows:

$$x_i = \left(x_i^1, \cdots, x_i^j, \cdots, x_i^n\right), y_i \in R;$$

x_i^j: the j-th feature of the i-th sample

The regression tree T (x; θ) is constructed as follows:

2.3 Random Forest

Random Forest by L.Brieman proposed, is a powerful performance of the multi-purpose classification and regression algorithm. Random Forest is composed of multiple random trees, and the average value of output of the random trees is used as the prediction result. The random tree is a variant of the decision tree, that is, in the decision tree construction process, the introduction of random nature: selecting k features from all features randomly, as feature set in the decision tree. And

then select an optimal feature from this subset for partitioning. Specifically, the range of values in step 2 of algorithm 1 becomes a randomly selected subset of attribute sets A. In addition, each time the sample set used to train the random tree needs to be sampled from the original data set D by self-sampling. Random Forest algorithm proposed that the size of the selected attribute subset is log_2 n, and the number of samples collected by self-sampling method is |D| [11].

Random Forest is constructed as follows:

3. Bus arrival time prediction model

3.1 Feature selection

Bus arrival time depends mainly on the traffic conditions and the length of the traffic section, and the traffic conditions depend on time, weather and other factors. This paper analyzes traffic data of Beijing 607 buses, combined with the actual traffic situation in Beijing, and extracts the traffic junction number, travel distance, date type, time period, precipitation, visibility of six influencing factors:

1: traffic junction number, travel distance

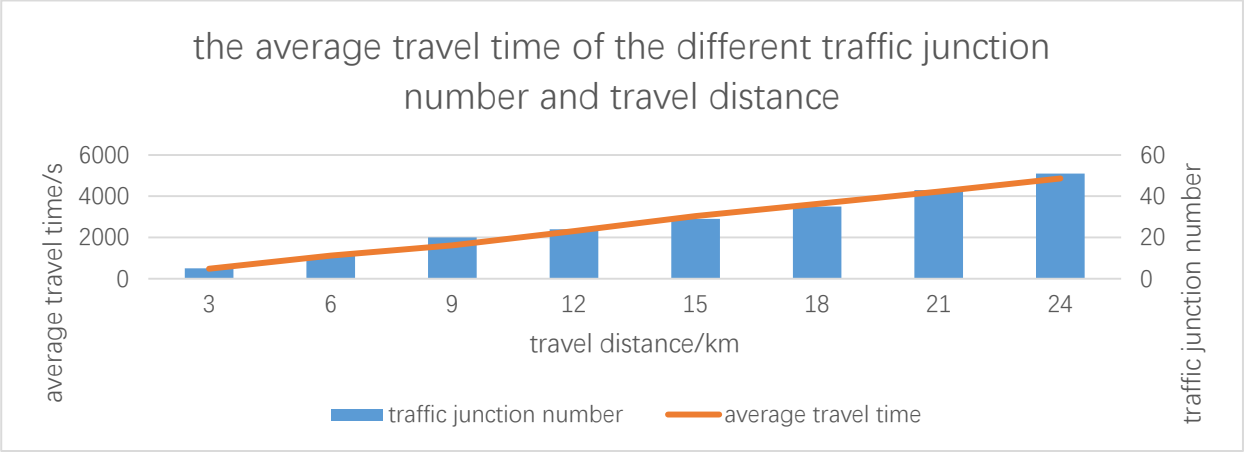

Fig. 1 the average travel time of the different traffic junction number and travel distance

This paper analyzes traveling data of Beijing 607 buses, and extracts the relationship between the number of traffic junctions, travel distance and the average travel time as showed above. The statistical results show that the average travel time of the bus increases as the travel distance and the number of traffic junctions increases.

2: date type

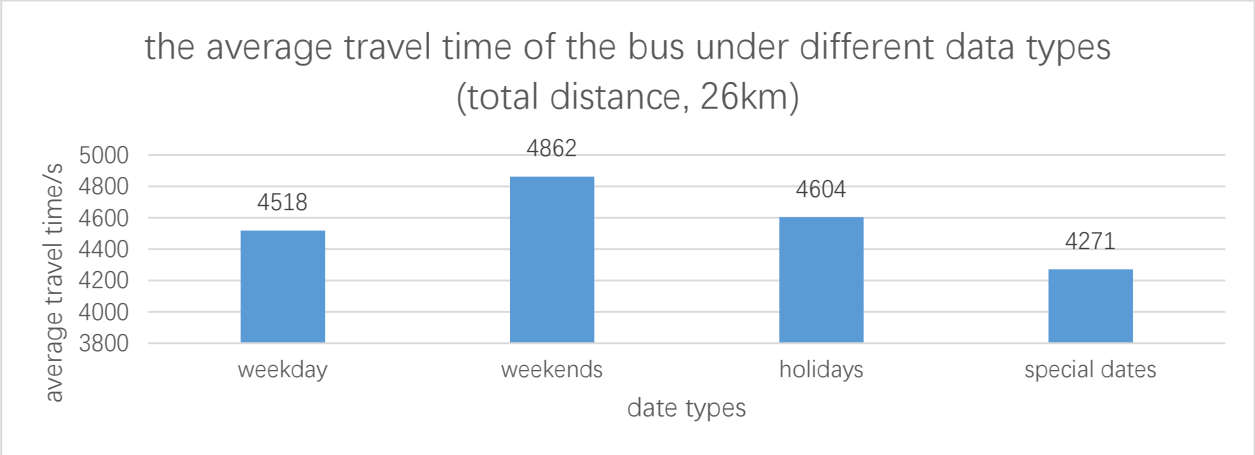

Fig. 2 the average travel time of the bus under different data types

Due to the special nature of traffic conditions in Beijing (odd-even license plate policy), which led to traffic conditions at working day and holiday conditions are different. In this paper, the date type is extracted as one of the influencing factors, and the date type is divided into four categories: weekday, weekends, holidays and special dates. Special dates represent the smaller traffic volumes such as the date when odd-even license plate policy effect. The results demonstrate that the travel time is shorter

due to the small traffic volume, traffic pressure and excellent driving conditions. And because the citizens going out at weekend frequently, weekend Beijing vehicles are not limited to the reasons, resulting in a longer travel time.

3: time period

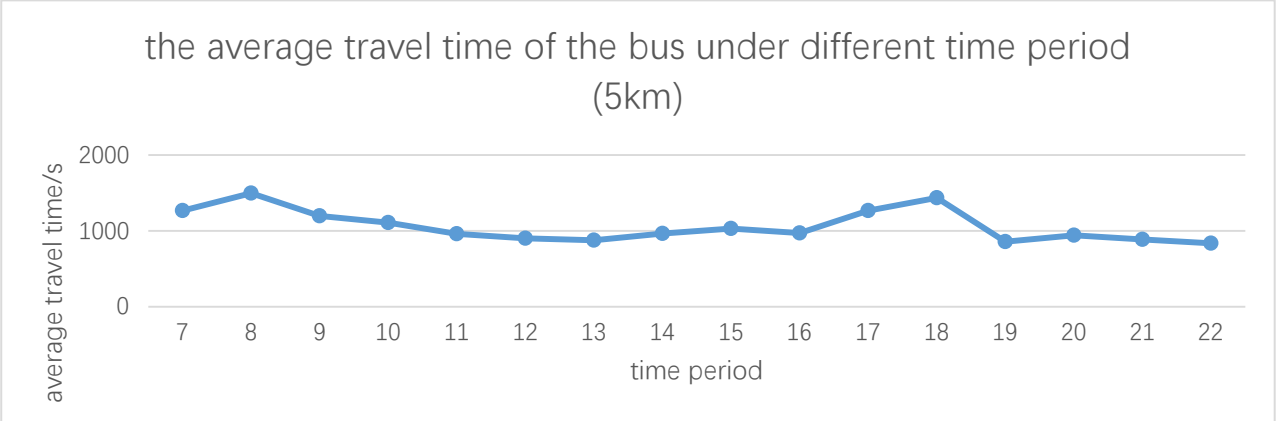

Fig. 3 the average travel time of the bus under different time period

Peak hours of work and leisure time is very different, as shown in Figure, average travel time increase when buses at 7:00 to 10:00 and 16:00 to 19:00. Therefore time period is the impact of traffic conditions one.

4: precipitation, visibility

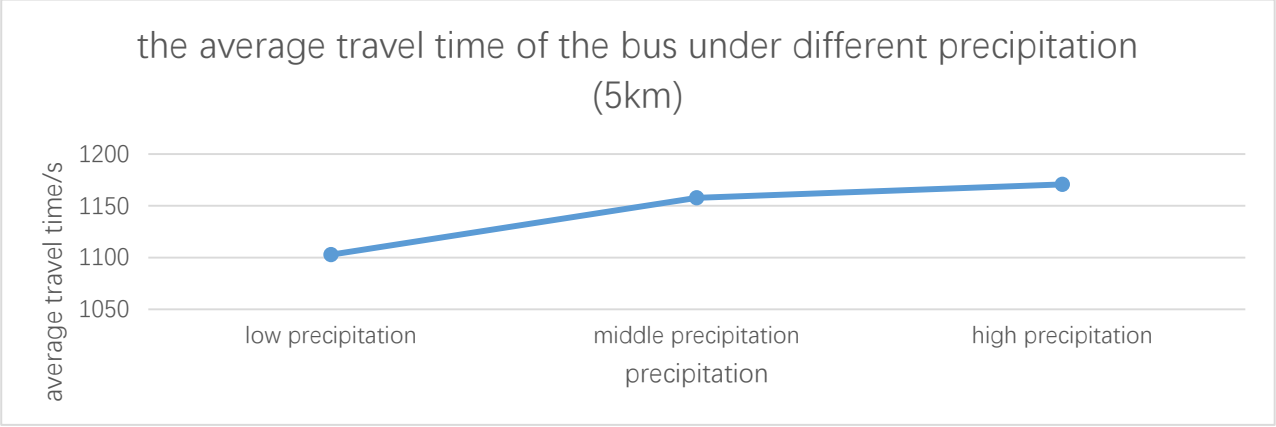

Fig. 4 the average travel time of the bus under different precipitation

The weather is also affecting the traffic situation. In order to quantify the weather condition, this paper uses the precipitation and visibility to do measurement.

3.2 Application of Random Forest

The regression tree is the basic composition of the random forest, and has a lot of forms on the choice of the loss function. Since this paper address the issue of the arrival time estimation problem, that is, the regression problem, this paper chooses the square loss function which is suitable for the regression problem. In addition, the number of regression trees in the random forest also affects the generalization ability of the algorithm. In this paper, 5 kinds of random forest models were constructed. Each containing 50, 100, 200, 400 and 800 decision trees.

The input of the random forest model is composed of six kinds of traffic junction number(x^1), travel distance(x^2), date type(x^3), time period(x^4), precipitation(x^5), visibility(x^6). The output of the model is the predicted value of the travel time of the bus on the route.

4. Numerical test

This paper collected more than 80,000 travel data of Beijing 607 buses a year, and randomly selected more than 60,000 data as training data to construct random forest model, and the remaining 20000 pieces were used as test data. The Mean Absolute Percentage Error (MAPE) was used as the measure to compare the error of each models. And the random forest model was compared with the

historical average model, the neural network model and the support vector regression machine. The results are displayed in the table below.

$$\text{MAPE} = \frac{1}{m}\sum_{i=1}^{m}\frac{|\widehat{y_i} - y_i|}{y_i} * 100\%;\ m\text{: the number of data}$$

y_i: the actual value of bus travel time for ith data

$\widehat{y_i}$: the predict value of bus travel time for ith data

Table 1 Comparison of predictive error of the random forest models

Number of decision trees	MAPE of Using GPS Geodetic Coordinate System	MAPE of Using Space Rectangular Coordinate System
50	24.31%	23.89%
100	23.24%	22.37%
200	21.17%	20.74%
400	20.85%	20.61%
800	20.73%	20.43%

Table 2 Comparison of predictive error of different models

Models	MAPE
historical average model	23.89%
neural network model	22.37%
decision tree model	20.74%
support vector regression machine	20.61%
random forest model (using Space Rectangular Coordinate System, number of decision trees=200)	20.43%

Table 1 records the impact of the number of decision trees in the random forest model on the predictive effect. The experimental results show that with the increase of the number of decision trees, the final prediction error of random forest shows a decreasing trend. Taking Space Rectangular Coordinate System as the data preprocessing condition as an example, the prediction error decreases by 0.87% on the basis of doubling the number of decision trees each time. When the number of decision trees is increased from 100 to 200, the prediction error decreases by 1.64%. But when the number of decision trees reaches a certain degree (the experiment is 200 decision trees), the error will decrease slowly, and the number of large decision trees will increase the training task. Although the distributed computing can be utilized to speed up the training, but also occupy the computing resources, the actual application of the number of decision trees to take a compromise approach. The experimental results also show that compared with using Geodetic Coordinate System as base of data preprocess, using Space Rectangular Coordinate System can reduce the error, about 0.454% on average. And maximum error reduction is 0.87% when the number of decision trees is 200. It is proved that the predictive effect of data preprocessing using Space Rectangular Coordinate System is better than that using the original GPS Geodetic Coordinate System.

Table 2 records the error of the bus arrival time prediction with different algorithm models. In this comparison, this paper selected the random forest model with using Space Rectangular Coordinate System for data preprocessing and the number of decision trees was set to 800. The experimental results demonstrate that the prediction error of the random forest model is 4.4% lower than the average historical model, and lower than the prediction error of the support vector regression machine by 1.24%. At the same time, it can be seen from the nature of the random forest algorithm that the basic unit (decision tree) is independent of each other, so the use of distributed computing method to train the random forest model will not affect the model results, in the training process is superior to

the neural network or other algorithms that use gradient descent as the basic solution. At the same time, because the basic idea of the basic unit (decision tree) training is to divide the value of the feature, Random Forest is more suitable for modeling discrete data such as date type compared with other algorithms.

5. Conclusions

Accurate prediction of bus arrival time depends on modeling the traffic environment. This paper analyzes the driving data of 607 buses in Beijing for one year, preprocesses the data under the condition of Space Rectangular Coordinate System, and extracts six influencing factors such as traffic junction number, travel distance, date type, time period, precipitation, visibility. Using Random Forest to construct the bus arrival time prediction model with a low error rate. In this paper, the arrival time prediction scheme is optimized from the aspects of narrowing the error. For other requirements of the arrival time prediction system, such as real-time and extensibility, did not do more research. And a more comprehensive study is required in the future to provide a complete system for the arrival time prediction system.

All manuscripts must be in English, also the table and figure texts, otherwise we cannot publish your paper. Please keep a second copy of your manuscript in your office. When receiving the paper, we assume that the corresponding authors grant us the copyright to use the paper for the book or journal in question. Should authors use tables or figures from other Publications, they must ask the corresponding publishers to grant them the right to publish this material in their paper.

References

[1] Shao Yuan, Song Jiahua. Traffic Congestion Management Strategies and Methods in Large Metropolitan Area. Urban Transport of China. 2010, 08(6).

[2] Sun Huijun, Si Bingfeng, Wu Jianjun. Combined Model for Flow Assignment and Mode Split in Two-modes Traffic Network. Journal of Transportation Systems Engineering and Information Technology. 2008, 8(4).

[3] Zhang M, Xiao F, Chen D. Bus Arrival Time Prediction Based On Gps Data. International Conference on Transportation Engineering. 2015:1470-1475.

[4] Chen Guojun, Yang Xiaoguang, Zhang Dong, Teng Jing. Historical Travel Time Based Bus-Arrival-Time Prediction Model. Internation Conference of Chinese Transportation Professionals. 2011:1493-1504.

[5] Gong J, Liu M, Zhang S. Hybrid dynamic prediction model of bus arrival time based on weighted of historical and real-time GPS Data. Control & Decision Conference. 2013:972-976.

[6] Ranhee Jeong, Laurence R Rilett. Prediction Model of Bus Arrival Time for Real-Time Applications. Annual Meeting of the Transportation-Research-Board. 2005:195-204.

[7] Jian Pan, Xiuting Dai, Xiaoqi Xu, Yanjun Li. A Self-learning algorithm for predicting bus arrival time based on historical data model. Cloud Computing and Intelligent Systems (CCIS). 2012:1112-1116.

[8] Chien I J, Ding Y, Wei C. Dynamic Bus Arrival Time Prediction with Artificial Neural Networks. Journal of Transportation Engineering. 2014, 128(5):429-438.

[9] Yu Bin, Yang Zhongzhen, Lin Jianyi. Bus Arrival Time Prediction Using Support Vector Machines. Journal of Intelligent Transportation Systems. 2007,10(4).

[10] J.R. QUINLAN. Induction of Decision Trees. Machine Learning. 1: 81-106, 1986.

[11] LEO BREIMAN. Random Forests. Machine Learning. 45, 5–32, 2001.