Tandem Prediction Research on LSTM-DT

Min Wang¹*, Xin Liu²
¹the Department of Basic Education, Shandong Jiaotong University, Weihai, China
²Weihai Institute of Marine Information Science and technology, Shandong Jiaotong University, Weihai, China

*Corresponding author’s email: 224028@sdjtu.edu.cn

Abstract. In order to effectively avoid the problem of bus tandem, the factors and prediction model of bus tandem based on multi-source data are studied. Due to the problems of complex attributes and low prediction accuracy, this paper presents a prediction method based on long short time memory and decision tree (LSTM-DT). Based on the existing historical GPS data, a variety of input data schemes are formed by different combinations of static spatial data and dynamic spatial data. In the process of tandem prediction, the long short time memory (LSTM) model is used to predict the time series characteristics of stations with different intervals, and the attribute and weight are adjusted by combining with decision tree. Through the LSTM-DT model, it gives the comparison of the tandem prediction in different time periods and different station intervals. The experimental results show that the LSTM-DT prediction model has better generalization ability than the traditional model in the case of more input factors, and has higher prediction accuracy for the tandem at a certain interval.

Keywords. Intelligent transportation; tandem; LSTM; prediction; decision-tree

1. Introduction
Accurate bus tandem prediction can better dispatch vehicles, schedule trips, reduce traffic accidents, and shrink government subsidies; at the same time, it can improve the attractiveness and service level of buses, increase the core competitiveness of buses, and reduce environmental pollution. Therefore, in-depth research on bus stringing prediction has been carried out at home and abroad. With the continuous development of big data and artificial intelligence, the collection and processing of bus operation data is more convenient, which promotes the feasibility and accuracy of bus tandem prediction to improve continuously.

Gong et al [1] assign different weights to historical and real-time data to ensure the real-time and accurate prediction time, and achieve the prediction based on the real-time location of the vehicle. In this model, the differential GPS data is used as input to obtain the stopping time of stations and the average travel time of two stations, taking into account the route, time and time period. Chen et al [2] built a prediction model based on Link and Section respectively based on the data obtained from the Automatic Vehicle Location System (AVLS), taking into account the following factors: time, time period and road section. It is verified that when it is not a working day, there is no difference between the two prediction results, when it is a working day, the Section-based model has better prediction.
effect; when the road traffic conditions are better, the same Section-based model has better prediction effect; when the vehicle enters the station, the prediction accuracy decreases, the Section-based model still has better results than the Link-based prediction model. Pang Junbiao et al [3] used a classical arrival prediction method for experimental comparison based on the largest existing bus operation database in Beijing, where different road situations and possible traffic conditions are included, and concluded that the regression theory-based method is more stable, in which the Linear Regression on Trajectories (LRT) [4] The algorithm is optimal. Yu et al. [5] established a multi-line SVM prediction model by inputting the headway, run-time weighted values and run-time weighted averages of different vehicles on different lines into the model, and the experimental results also showed an improvement in accuracy. Similarly, Yu et al [6] compared the accuracy of prediction with each other by adding the forgetting factor, Grubbs test to remove anomalous data and simultaneous operation, respectively. For the high-dimensional data problem, SVM can also be effectively solved with limited samples, with some adaptability, but the randomness of the real-time string bus features leads to low prediction accuracy. Therefore, in this paper, based on the collection method and data format of public transportation data, pre-processing such as interpolation and normalization is performed on the collected data, and prediction data samples are identified based on the operating characteristics and influencing factors of public transportation vehicles. On this basis, in order to solve the multi-factor problem that cannot be solved by a single model, a combined model LSTM-DT is given for the string vehicle prediction model, which accurately reflects the temporal and spatial relationships of the complex characteristics of string vehicles and improves the prediction accuracy.

2. Data acquisition and processing

2.1. Data acquisition
The data is obtained from the operation command and dispatch system of the bus group, and is transmitted in real time from the GPS device and GIS monitoring to the short message server using GPRS network, and then to the monitoring center via DNN dedicated line, thus monitoring and dispatching the vehicle status in real time. The collected data including: vehicle number, route, running time, running position (latitude and longitude), running speed, etc., are transmitted to the monitoring center according to a specific transmission protocol, and the storage format is TXT form. The experimental data are collected based on the vehicle GPS device, mainly including historical data and real-time data, which are imported into the database for easy data analysis and processing.

2.2. Data Process
The GPS device is collected every 3s, and the average speed of the bus is 20-30km/m. For the measurement error generated during the collection process, the missing data values are interpolated to ensure that the vehicle location information can be obtained every 1s. In this paper, three sample strips interpolation method is used to solve the shortcomings caused by the increasing number of nodes is not smooth, after the known need to interpolate the bus GPS data, most of the polynomial interpolation will be determined, can not be locally adjusted, three sample strips method according to the value of the function to calculate the derivative value, and then use the segment Hermite interpolation to ensure the smoothness of the interpolation point. Using \( \{ T_0, T_1, ..., T_i, ..., T_k \} \) to denote the collected data points whose coordinate position is \( [x_i, y_i] \), then set the GPS data of each bus as shown in equation (1).

\[
\begin{align*}
X &= [x_0, x_1, ..., x_k] \\
Y &= [y_0, y_1, ..., y_k]
\end{align*}
\]

The interpolation function in each segment is shown in equation (2).

\[
f_i(x) = a_{i3}x^3 + a_{i2}x^2 + a_{i1}x + a_{i0}
\]
3.1. Basic Principle
The LSTM model is a deformation of the RNN [7] network, and the overall structure of the LSTM model has not changed much compared to the standard RNN. The improvement of the LSTM [8] network for the RNN is the hidden layer, with a unique memory and forgetting mode. The network structure is selected using activation function, loss function and optimizer function, and the LSTM network model with different layers such as single layer and double layer are experimented separately to decide the number of network layers of LSTM according to the prediction results. And on this basis, the LSTM-DT model with attribute setting and weight adjustment based on decision tree pruning is given, considering the changes of prediction aging, temporal factors and spatially related attributes in different time periods.

The model is based on historical GPS data and decomposes other properties in the spatial domain from a time domain perspective into two parts: temporal autocorrelation properties and spatial cross-correlation properties. According to the data analysis, the traffic flow data has a certain pattern in a particular location or region over a certain period of time. Between the traffic data of the same weekday, the same rest day or different periods of the same day generally have the same pattern, therefore, according to the hidden cycle change pattern by DT to do attribute pruning and weighting adjustment helps to get more accurate prediction results.

3.2. Model Assumption
The model assumes that the vehicle’s current station m, steps (s=1,2,…,n) , the output value of m + stime series for the construction of the LSTM-DT model. the LSTM layer selection of the default tanh function as the activation function, loss function (loss) as the error loss, decision tree as an optimization algorithm, after experiments compared to the traditional algorithm The equation of state for LSTM-DT is shown in equation (3).

\[ STAT^m(t) = T^m_{kd} = f^1(t_{kd}^m+s) \]  

where \( STAT^m(t) \) is the time at the current station, \( T^m_{kd} \) is the observation coefficient, and \( V^m(t) \) is the observation noise. For the \( k \)th trip of a vehicle, the data \( t_{kd}^m+s \) at the \( mth \) + \( i \)th station.

Road congestion index evaluation is determined by saturation, which is an important indicator of the level of service on the road network. Road traffic saturation is the ratio of road traffic volume to the design capacity of the road. The higher the saturation value, the more congested the road is. Since the level of road service and congestion is limited by many factors, it is difficult to consider multiple factors in practice, and the saturation value is often used as the main indicator for evaluating the level of service. The maximum capacity of a road is calculated by equation (5), where \( v \) represents the driving speed \( (km/h) \); \( t_0 \) is the maximum hourly headway \( (s) \); \( l_0 \) is the minimum headway interval \( (m) \).

\[ N_{max} = \frac{3600}{t_0} = \frac{3600}{l_0/(v/3.6)} = \frac{1000v}{l_0} \text { (vehicle/h)} \]  

During the forward prediction calculation, the output of the implicit layer jointly enters the forgetting gate, the value of which at moment \( t \) is shown in Eq. (6).

\[ f(t) = \sigma(w_k[h(t),x(t+1),S(t),D(t)] + z_k) \]  

\( f(t) \) is the state value of the oblivion gate at time \( t \); \( \sigma \) is the sigmoid function of the oblivion gate; \( w_k \) is the weight matrix; \( h(t) \) is the output data at time \( t \); \( x(t+1) \), i.e., \( t_{kd}^m+s \) is the time series of stations \( m+s \); \( S(t) \) is static data, including time period \( T \), station type \( P \), weather \( A \), intersection \( X \), and \( D(t) \) is dynamic spatial attribute data ; \( z_k \) is the bias matrix. In order to improve the rationality of the attribute selection in the crosstalk prediction, the statistical analysis of different temporal and
spatial attribute correlations is performed using Pearson's correlation coefficient, which is used as the basis for decision tree pruning.

The output gates control the output information, the state of which is shown in Eq. (7) and Eq. (8).

\[ p(t) = \sigma(w_p, [h(t), x(t+1), S(t), D(t)] + z_p) \quad (7) \]

\[ h(t) = p(t) \ast \tanh(C(t)) \quad (8) \]

\( p(t) \) represents the output cell state after input value \( x(t+1) \) through sigmoid function; \( h(t) \) is the output gate state obtained by multiplying input value \( x(t+1) \) through tanh function and \( p(t) \), i.e. t moment, the time series of site \( m \cdot s \) obtained after hidden layer, \( C(t) \) is the current state value of memory cell, the output of the cell depends on output gate \( h(t) \) and the current state value of memory cell.

The above is the forward computation process. The reverse computation process defines a centerloss loss function, based on the mean squared error loss of \texttt{mse} listed in the keras document, with the average of all the features of the sample as the class center \( c \), giving the centerloss function as shown in equation (9).

\[ l(t) = ||h(t) - \bar{x}(t)||^2 \quad (9) \]

where \( h(t) \) is the output information and \( \bar{x}(t) \) is the average of the time samples of stations in the line. The goal is to minimize the loss function over the entire time series. Considering the historical and future spatial information of the time series at the same time helps to construct a complete sequence model, so the characteristic representations of the clockwise and counterclockwise arrival times are obtained by two-way LSTM network extraction, processing the arrival time series from the going and coming directions, respectively, and the resulting clockwise arrival time is denoted \( T_g \) and coming arrival time is denoted \( T_c \). The time series represents \( T_g \) by \( T_g \) and \( T_c \). The corresponding time is averaged as shown in Eq. (10), and finally the stringing probability is obtained by comparing the arrival times.

\[ T_a = AVERAGE(T_g, T_c) \quad (10) \]

4. LSTM-DT prediction algorithm

4.1. Algorithm Description

The overall flow of the LSTM-DT-based crosstalk prediction is shown schematically in Figure 1, and the following processes need to be completed in the model.

1) Data division and pre-processing: divide the data into two parts, training set and test set, define the input variables of LSTM network, and solve the data to complete the formatting of the data obtained.
2) DT-based attribute pruning and weight adjustment: in the problem of predicting crosstalk, the adjustment of weights is complicated when there are too many attributes, the selection and range settings of various attributes are difficult to determine in the LSTM model, and the DT algorithm is easy to determine the influence of attributes on the prediction, therefore, an implicit layer is added in the middle of different time periods, and some attributes are removed by pruning from the decision tree. The state of the implicit layer added between the input layers of different time periods is \( \{p_1, p_2, ..., p_t\} \), and the implicit information of each time period is weighted by the decision tree to determine its influence on the prediction results. Equation (11) for the implicit attributes is shown below.

\[
P_t = \sigma(\lambda_{t-1} + B_\lambda t + C D_t)
\]

where \( \sigma \) is the nonlinear unit, \( \lambda \) represents the weight of historical GPS data on the current input, \( B \) represents the weight occupied by static data, \( D_t \) represents the static attribute at moment \( t \), and \( D_t \) represents the dynamic attribute at moment \( t \). The LSTM model is based on the LSTM-DT model.

3) LSTM-DT based timing prediction: the approximate range of each parameter in the LSTM is determined empirically to construct N LSTM models. Input the training set into the LSTM model with different parameters, compare the output of the predicted arrival time of the validation site with the actual value, and complete the training of each LSTM model through the loss function and optimization function. Then the test set data is input into the trained N LSTM models, and the N predicted vehicle arrival time series of the desired site is obtained.

4) Finally, through the obtained time series, the string vehicle predictions for the same and different lines at the same site are performed.

### 4.2. Algorithm Improvement

The decision tree is used to dynamically adjust the input information for different time and spatial segments in the LSTM algorithm. Then the input gate is used to determine how much new information is needed to enter the cell state, and finally the cell state is filtered and output through the output gate. The information in the Oblivion Gate is determined by the "tree pruning" of the decision tree.

The decision tree uses a top-down recursive classification method to construct a decision tree classifier by "tree generation" and "tree pruning". The decision tree algorithm is built in the following steps.

**Step 1:** Given a list of training sample attributes and an empty decision tree.

**Step 2:** determine if all nodes are traversed end, if "yes" then the algorithm is terminated, "no" to determine if all samples in the sample set \( T \) is the same category, "yes" then store attribute \( X \) for pruning nodes, and leaf node generation; "no" then go to step Step 3.

**Step 3:** adopt a method for selecting test attribute \( X \) and enable the generation of test nodes.

**Step 4:** if the features of \( X \) are denoted by \( x_1, x_2, ..., x_n \) representation, \( T \) is partitioned according to the different \( x \) features: \( n \) subsets \( T_1, T_2, ..., T_n \).

**Step 5:** removing attribute \( X \) from the attribute table.

Step 6: go to Step 2 until a decision tree is generated that correctly classifies the training sample.

### 5. Simulation experiments and results analysis

#### 5.1. Data Sets

This experiment uses the October 2020 data of the Rongcheng 107 bus. the AVL data processing process removes records that are not within the variation of the latitude and longitude of the route, as shown in Table 1.

| VehicleId | Time | Longitude | Latitude | Speed | Azimuth |
|-----------|------|-----------|----------|-------|---------|

5
After the raw vehicle location data is matched to a station, the sequence of stations matched to successive sampled records of the same vehicle should correspond to the line's station layout, while the inbound state alternates between in-station and out-station. The sampling time of the 1st record of the selected inbound state from outstation to station is the inbound time; the sampling time of the 1st record from station to outstation is the outbound time. The inbound and outbound time information of the vehicle is obtained, as shown in Table 2.

**Table 2** Arrival and Departure Time

| RecordId | VehicleId | StopId | StopName          | ArrivalTime       | DepartureTime     | TripId |
|----------|-----------|--------|-------------------|-------------------|-------------------|--------|
| 1        | 107       | 1      | LiangWanCheng     | 2020/10/08 7:31:08 | 2020/10/08 7:31:48 | 122.530 938 37.168 889 2 56 |
| 2        | 107       | 2      | HaLiGong          | 2020/10/08 7:31:48 | 2020/10/08 7:32:05 | 122.518 325 37.169 320 32 42 |
| 3        | 107       | 3      | MaJiaZhuang       | 2020/10/08 7:32:05 | 2020/10/08 7:32:46 | 122.513 439 37.169 479 16 42 |
| 4        | 107       | 4      | YueHaiYuan        | 2020/10/08 7:32:46 | 2020/10/08 7:33:02 | 122.508 421 37.168 386 29 33 |
| 5        | 107       | 5      | YueHaiYuan        | 2020/10/08 7:33:02 | 2020/10/08 7:33:42 | 122.497 718 37.164 755 0 38 |
| 6        | 107       | 6      | YueHaiYuan        | 2020/10/08 7:33:42 | 2020/10/08 7:34:01 | 122.495 454 37.165 395 0 29 |

GPS data is not continuous and reliable due to the inherent uncertainties of various devices and networks. In addition to cases where two adjacent samples happen to miss certain stations, there are also cases where GPS data is lost or drifted. After the direction of travel is determined, the inbound and outbound times of the missing stations need to be made up. The arrival time prediction of spatial data based on LSTM-DT is compared horizontally with traditional common prediction algorithms SVN and KNN for four dimensions: model prediction accuracy, model prediction time, model generalization ability, and cross-site prediction performance. In order to verify the generalization ability of the model, the data are divided into three types of input, including historical GPS data, static spatial data and dynamic spatial data. The static spatial data include time period, station type, weather and intersection, while the dynamic spatial data include green signal ratio, road congestion level and number of card users.

5.2. Results Analysis

![Figure 2. SVN、KNN、LSTM-DT prediction](image-url)

Different prediction models using the same dataset, the dataset is historical data, static spatial data and dynamic spatial data. Three prediction methods, SVN, KNN, and LSTM-DT, are used to perform crosstalk prediction, and their generalization ability and prediction error are evaluated by comparing the predicted data with the original data. The results of the statistical data are shown in Figure 2.
For different inputs, the output predictions are shown in Figure 2, and the time model of the LSTM-DT network algorithm fits better to the original data when the input data latitude increases compared to the traditional model. From Figure 3, it can be seen that the LSTM-DT algorithm is less efficient in execution compared to the traditional algorithm in the off-peak time period.

**Figure 3.** prediction time

To test the generalization ability of the model, the 3681 valid data from 107 routes from October 8 to October 19 are selected, of which 2678 data are used as training samples to learn from different models, and the remaining 2678 data are used as training samples. Data is tested as new data. The degree of agreement between the predicted value and the true value is measured by the loss function. The error value (loss) statistics of the optimal model after updating the weights and biases of the model after the prediction error correction were performed, as shown in Figure 4, the loss values of each model converged to a value below 0.19, all with good prediction effect.

**Figure 4.** training error

In order to further statistically model the performance of the model on the test set, the prediction problem for multi-site span was tested, as shown in Figure 5, although the prediction duration of LSTM-DT is yet to be optimized, the prediction results of LSTM-DT are closer to the real data, reflecting its advantage in prediction accuracy for complex situations.

**Figure 5.** the comparison of prediction time

6. Conclusion
By combing the data of static and dynamic factors in bus bunching prediction, and analyzing the LSTM-DT algorithm in bus bunching prediction according to different combinations of input data, the
prediction accuracy, prediction time, generalization ability and cross-site prediction accuracy of the LSTM-DT algorithm relative to the traditional prediction algorithm, the experimental results show that, for complex cases, the prediction model based on the LSTM-DT algorithm in the input data dimension. The increased case has advantages in prediction accuracy and generalization ability, but the prediction time is relatively long. Optimizing the duration is a follow-up research effort.

7. Acknowledgment
This research was supported by Shandong Province science and technology plan project under Grant No.J17KB37.

8. References
[1] Gong J, Liu M, Zhang S 2013 Hybrid dynamic prediction model of bus arrival time based on weighted of historical and real-time GPS data// Control & Decision Conference. IEEE.
[2] Chen G, Yang X, An J, et al 2012 Bus-Arrival-Time Prediction Models: Link-Based and Section-Based. Journal of Transportation Engineering, 138(1):60-66.
[3] Pang J B, Hu J J, Huang J 2019 Comparison of bus arrival time prediction methods based on large scale data. Computer technology and development, (04): 1-7.
[4] Sinn M , Yoon J W, Calabrese F, et al 2012 Predicting arrival times of buses using real-time GPS measurements// International IEEE Conference on Intelligent Transportation Systems. IEEE.
[5] Yu B, Ye T, Tian X M, et al. 2014 Bus Travel-Time Prediction with a Forgetting Factor. Journal of Computing in Civil Engineering, 28(3):06014002.
[6] Yu B, Lam W H K, Tam M L 2011 Bus arrival time prediction at bus stop with multiple routes. Transportation Research Part C Emerging Technologies, 19(6):1157-1170.
[7] Kumar J , Goomer R, Singh A K 2018 Long Short Term Memory Recurrent Neural Network (LSTM-RNN) Based Workload Forecasting Model For Cloud Datacenters. Procedia Computerence, 125(01):676-682.
[8] Ma X, Tao Z, Wang Y, et al 2015 Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. Transportation Research Part C, 54(01):187-197.