IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 14, 2020, accepted February 1, 2020, date of publication February 5, 2020, date of current version February 17, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2971771

Passenger Flow Forecast of Rail Station Based on Multi-Source Data and Long Short Term Memory Network

ZHE ZHANG[1], CHENG WANG[1], YUEER GAO[2], YEWANG CHEN[1], AND JIANWEI CHEN[3]

[1]College of Computer Science and Technology, Huaqiao University, Xiamen 361021, China
[2]College of Architecture, Huaqiao University, Xiamen 361021, China
[3]Department of Mathematics and Statistics, San Diego State University, San Diego, CA 92182, USA

Corresponding author: Cheng Wang (wangcheng@hqu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China Youth Fund under Grant 51608209, in part by the Project of Natural Science Foundation of Fujian Province of China under Grant 2017J01090, in part by the Project of Quanzhou City Science and Technology Program of China under Grant 2018Z008, in part by the 2018 Huaqiao University Research and Establishment of Postgraduate Education and Teaching Reform Project under Grant 18YJG28, and in part by the Huaqiao University Postgraduate Research Innovation Ability Cultivation Program under Grant 18014083027.

ABSTRACT The existing rail station passenger flow prediction models are inefficient, due to that most of them use single-source data to predict. In this paper, a novel method is proposed based on multi-layer LSTM, which integrates multi-source traffic data and multi-techniques (including feature selection based on Spearman correlation and time feature clustering), to improve the performance of predicting passenger flow. The experimental results show that the multi-source data and the techniques integrated in the model are helpful, and the proposed method obtains a higher prediction accuracy which outperforms other methods (e.g. SARIMA, SVR and BP network) greatly.

INDEX TERMS Rail transit passenger flow, prediction model, long short term memory network, multi-source data, spearman correlation, K-means.

I. INTRODUCTION

With the rapid development of urban rail transit and the continuous improvement of information management system, a large number of passenger travel data have been generated. How to accurately estimate the passenger flow of the rail station has become a research hotspot in the scientific community. It is well known that accurately prediction on the passenger flow could not only help us control the situation of passenger transportation, but also is useful for making a reasonable plan for potential emergencies, which improves the city's emergency response capacity.

However, there are some problems in passenger flow forecasting currently, such as single data source, incomplete consideration of influencing factors, which yield low accuracy of existing methods, and seriously defect the management of the urban traffic.

In this paper, based on multi-source traffic data, a novel method, which take both of temporal and spatial factors into consideration, is proposed to predict passenger flow of rail station. Especially, it utilizes clustering algorithm in classifying features of temporal factor, due to that clustering algorithm, such as k-means, DBSCAN [1], [2], Density Peak [3], [4] etc., is an effective way to classify data into different categories automatically, and is quite suitable and applicable in this field. Furthermore, due to the time series characteristics of passenger flow, based on the advantages of LSTM network [5] in modeling time series data, a multi-layer LSTM network passenger flow prediction model is proposed as well to predict the entrance passenger flow of rail station.

The associate editor coordinating the review of this manuscript and approving it for publication was Songwen Pei.

II. RELATED WORK

There are many factors that affect the passenger flow of rail station, which could roughly be classified into two categories: temporal factors and spatial factors.

(1) Temporal factors are the factors which affect the passenger flow in time dimension. Most scholars regard the historical passenger flow as an important influencing factor [6]. For example, Liu *et al.* [5] use the passenger flow before the

target period as a key feature to predict the passenger flow. Meanwhile, the passenger flow is closely related to ''weekly information''[7], i.e. ''What day is today?'' (e.g. Monday, Tuesday.....Sunday) and ''the time of the target period''. Lijuan Liu, et al. also predicted the passenger flow of bus rapid transit (BRT) stations in a similar way [8].

(2) Spatial factors are the factors that affect the passenger flow in the spatial dimension according to the spatial location relationship of the station. Tang *et al.* [9] considered the impact of passenger flow of other stations in the rail network space on the passenger flow of the predicted rail stations, which improving the prediction accuracy. Junwen Zhou considered the multi-source data, introduced the third-party consumption data around the rail station that may affect the passenger flow, to predict the passenger flow [10].

Most of the current researches are based on the temporal features and single rail transit data itself, such as ''weekly information'' and ''the time of the target period'', the historical passenger flow data of the station itself, and the impact of other rail station passenger flow in the rail network space. In terms of spatial factors, there are less researches on multi-source data [9], and less explorations on the use of multi-source traffic data. Furthermore, to our knowledge, at present time, there is no research that uses other public transport (e.g. bus rapid transit) data to study the inter-spatial-relationship among different kinds of transportation, which leads to the lack of comprehensive consideration of spatial influence factors in passenger flow prediction and yields low performance. As far as ''the time of the target period'' is considered, some scholars directly used it as input, which has poor generalization [8], while other scholars divided the two categories of peak and off-peak time artificially [11], which has strong randomness.

In terms of prediction method, scholars at home and abroad have done a lot of researches on passenger flow prediction. There are several methods as follows: (1) Linear prediction model, Milenković *et al.* [12] used autoregressive moving average model to predict the railway passenger flow. (2) Based on the method of machine learning, Sun *et al.* [13] used the wavelet SVM model to predict the Transfer passenger flow of Beijing Rail Transit. Some scholars predict passenger flow based on neural network [14]. (3) Combined forecasting model. Jiang *et al.* [15] established a combined model based on grey support vector machine to improve the prediction accuracy of rail transit passenger flow. The current research results are of great significance for passenger flow prediction, but there are some limitations [16]: the linear prediction model cannot reflect the nonlinear characteristics of passenger flow [17]; the prediction accuracy of machine learning model is not enough; the generalization ability of combined prediction model is poor [18].

Theoretical analysis and comparison of methods: There are various different factors that affect the passenger flow, as shown in Table 1 which shows that it is superior to use multi-source. However, most existing works only make use of single factor, instead of multi-source, to predict passenger flow, yielding accuracy is not enough [12], such as linear prediction model, nonlinear prediction model, simulation prediction model, machine learning model, etc.

TABLE 1. Data source comparison.

Data source	advantage	disadvantage
Rail transit card swipe data	Highly relevant and accurate	Single data source
Rail transit card swipe data and the third-party consumption data	Multi-source data, more diversified influencing factors	The third-party consumption data is not accurate enough. Moreover, it is not strongly correlated with the passenger flow to be predicted.
Multi-source traffic data (Rail transit card swipe data and BRT card swipe data)	Multi-source data, more diversified influencing factors. The data is accurate and has a strong correlation with the predicted passenger flow.	High data quality is required. And data processing is complicated.

TABLE 2. [16] Method comparison.

Method Name	advantage	disadvantage
Linear Prediction Model[17]	It can well express the linear characteristics of influencing factors of passenger flow.	Inherent characteristics of short-term passenger flow cannot be fully acquired.
Simulation Prediction Model[19]	Its the most accurate restoration of the true situation, with high accuracy.	Modeling costs are high.
Shallow Neural Network and Machine Learning Model[13]	The model has the advantages of simple structure, easy modeling and high accuracy.	It is easy to have problems of over-study or under-study, which can not reflect the characteristics of time series.
Deep Learning Lstm Model[9]	It can better learn the time series features of data and has high accuracy	Over fitting

The comparison of various existing methods is shown in Table 2 [16]. Compared with them, the LSTM network model is able to effectively extract features from the source data, which is difficult for the traditional linear model and other machine learning methods. Furthermore, its modeling cost is far lower than simulation prediction models [19], and is suitable for time series data.

III. NOTATIONS AND TASK DESCRIPTIONS

Before introducing the proposed method, some notations used in this paper are addressed as below.

Let $DAY = \{day_1, day_2, \ldots, day_e\}$ be a day set, and $Week(day_i)$ be the ''weekly information'' of day_i; Δt be a time interval (e.g., 0.5, 1 or 2 hours), and a day is divided into

q segments, where $q = 24/\Delta t$, which forms a time period vector $T = \{T_1, T_2, \ldots, T_q\}$.

Let $ST = \{st_1, st_2, \ldots . st_m\}$ be a rail station set, $RailADST_w = \{railadst_{w,1}, railadst_{w,2}, \ldots, railadst_{w,z}\}$ be the adjacent rail station of st_w where $RailADST_w \subseteq ST$, and $BrtADST_w = \{brtadst_{w,1}, brtadst_{w,2}, \ldots, brtadst_{w,s}\}$ be the adjacent BRT station of st_w; $pflow_{i,j}^{st}$ be the entrance passenger flow of station st at time T_i on day_j, we also use $PFlow_{i,j}^{RailAdST_w} = \{pflow_{i,j}^{RailAdST_{w,1}}, pflow_{i,j}^{RailAdST_{w,2}}, \ldots.\}$ to represent the passenger flow set of the adjacent rail stations of st_w, and similarly $PFlow_{i,j}^{BrtAdST_w} = \{pflow_{i,j}^{BrtAdST_{w,1}}, pflow_{i,j}^{BrtAdST_{w,2}}, \ldots.\}$ represent the passenger flow set of the adjacent BRT stations of st_w. Hence, $pflow_{i,j}^{RailAdST_{w,y}}$ is the passenger flow of the y^{th} adjacent rail station of st_w, and $pflow_{i,j}^{BrtAdST_{w,y}}$ is the passenger flow of the y^{th} adjacent BRT station of st_w.

Let $Te = \{te_1, te_2, \ldots . . te_i, \ldots . . . , te_n\}$ be the feature set of temporal-spatial influence factors, where te_i represents the i^{th} feature.

Let $P = \{x_1, x_2, \ldots, x_n\}$ be a dataset, and $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ be the sorted sequence of P, where $x_{(1)} < x_{(2)} <, \ldots, < x_{(n)}$, and let $ind(P, x_i)$ be the index of x_i in the sorted sequence of P.

Definition 1 (Order Difference): Let x_i and x_j be two elements of P, $|ind(P, x_i) - ind(P, x_j)|$ is called the Order Difference.

Definition 2 (passenger flow): The number of passengers.

Our task is to predict the entrance passenger flow of a rail station.

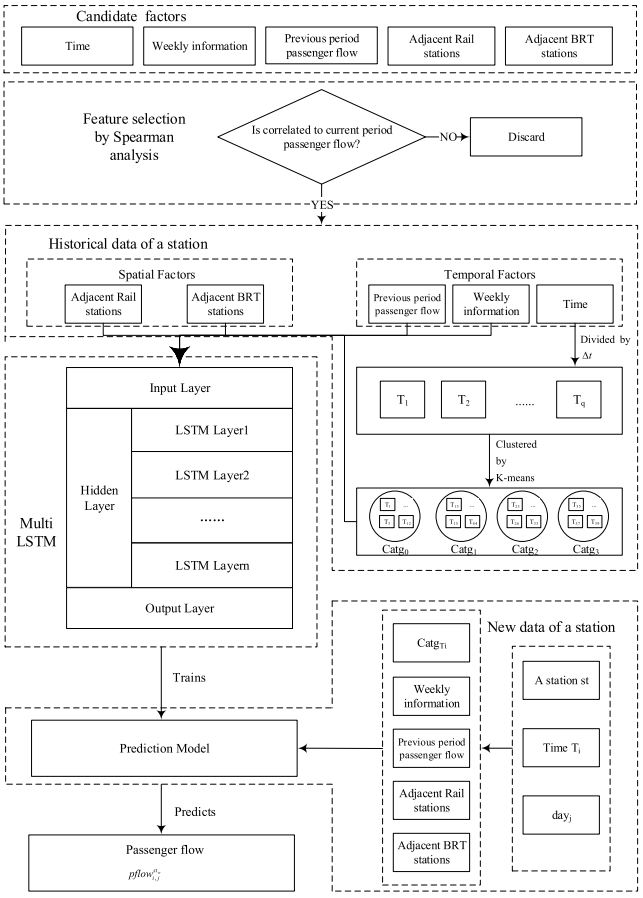

FIGURE 1. The framework of the proposed method.

IV. PASSENGER FLOW PREDICTION MODEL OF RAIL STATION BASED ON LSTM NETWORK

A. THE FRAMEWORK

The framework of the proposed method is shown in figure 1, by training the historical data it obtains a predictor which is used to forecast the passenger flow.

As we can see from the plot, firstly, we use Spearman algorithm to analyze the candidate influence factors, then we get the temporal and spatial factors that affect the passenger flow significantly. Secondly, we use k-means algorithm to cluster the feature of ''the time of the target period'' to get the category of the time which can simplify the proposed model. Thirdly, a multi-layer LSTM network passenger flow prediction model is established and trained by the historical data. Finally, the trained model is used to predict the entrance passenger flow of rail station in the actual situation.

B. TEMPORAL AND SPATIAL FACTORS

It is a common sense that time and space are the two most important factors which affect the passenger flow much. In this paper we analyze how important they are based on spearman correlation, and study on the mechanism of how they work, as below.

1) SPEARMAN CORRELATION

As mentioned above, the rail station passenger flow is affected by many factors that hide behind raw data. In order to find them, we use the spearman correlation analysis method, which is also known as ''rank correlation coefficient'', to test the correlation coefficient between the passenger flow and each factor.

Spearman correlation uses the appearance orders of two variables to analysis correlation and has no mandatory requirements for the distribution of original variables including normal distribution, which means that it belongs to nonparametric statistical method and has a wider range of application. Furthermore, it studies the correlation between two variables based on computing the order difference between them which is also called "rank difference method". In the case of no duplicate data, if one variable linearly depends on another one, the correlation coefficient is + 1 or - 1, and this is called the variable complete spearman order correlation.

2) TEMPORAL FACTORS

(1) As we know the passenger flow varies with time in a day, for example, Figure 2 plots the changes of passenger flow for different 4 days at *Lianban* rail station of *Xiamen* rail line

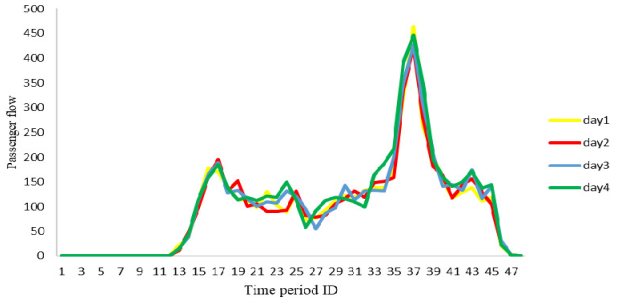

FIGURE 2. The changes of entrance passenger flow for different 4 days at *Lianban* rail station.

TABLE 3. The correlation coefficient values of influencing factors.

influencing factors	correlation coefficient
"the entrance passenger flow of *Hubindonglu* rail station in the previous period"	0.900
"the entrance passenger flow of *Liabhualukou* rail station in the previous period"	0.897
"the entrance passenger flow of BRT *Lianban* station in the previous period"	0.884
"the entrance passenger flow of *Lianban* rail station in the previous period"	0.909
"weekly information" (with daily entrance passenger flow)	0413
"the time of the target period"	0.622

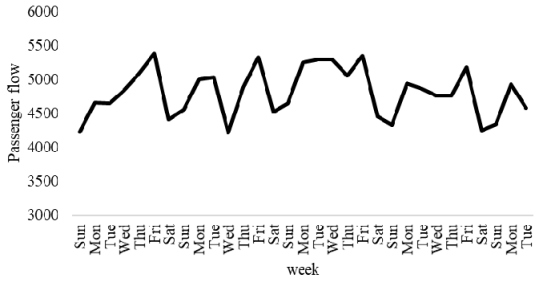

FIGURE 3. The change of daily entrance passenger flow at *Lianban* rail station in July 2018.

1 ($\Delta t = 0.5$ hour, then $q = 48$, $T = \{T_1, T_2, \ldots, T_{48}\}$). We can see that the four curves are similar, indicating that the passenger flow is roughly unimodal, where the main peak of passenger flow appears between $[T_{36}, T_{38}]$. Given a rail station st_w at time T_i on day_j, its passenger flow $pflow_{i,j}^{st_w}$ is affected by T_i. As shown in the last row of Table 3, the correlation coefficient between the target period passenger flow of *Lianban* rail station and "the time of the target period" is about 0.622.

(2) It is also well known that the daily passenger flow of the rail station changes with the "weekly information". Figure 3 plots the daily entrance passenger flow of *Lianban* rail station in July 2018. It is observed that the impact of "weekly information" is very significant, i.e., the distribution of the passenger flow varies weekly, that it is quite different on different day, and is similar on the same week day

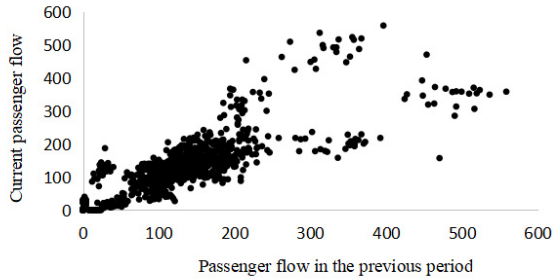

FIGURE 4. The correlation between the current period passenger flow and the previous passenger flow of *Lianban* rail station.

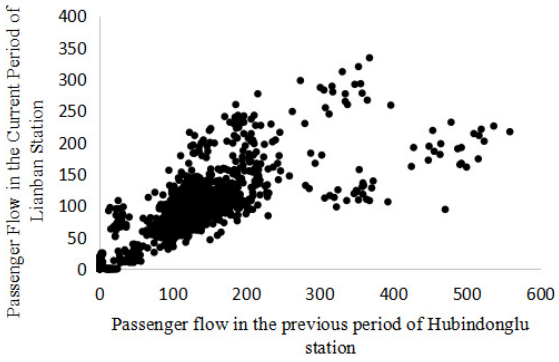

FIGURE 5. The correlation diagram between the current passenger flow of *Lianban* Station and the previous passenger flow of *Hubindonglu* rail station.

(e.g., the flow on Monday is similar to that of other Monday). As shown in the penultimate row of Table 3, the correlation coefficient between the daily passenger flow of *Lianban* rail station and "weekly information" is about 0.413.

(3) In addition, given a station st_w on day_j, its passenger flow of the current period is closely related to that of its previous period, i.e. $pflow_{i,j}^{st_w}$ is affected by $pflow_{i-1,j}^{st_w}$. Figure 4 shows the correlation between the current period passenger flow and the previous period passenger flow of *Lianban* station. As shown in Figure 4 that there is a strong correlation between them. So, the passenger flow in the previous period is also an important influence factor. As shown in the 4$^{\text{th}}$ row of Table 3, the correlation coefficient is about 0.909.

3) SPATIAL FACTORS

(1) Due to the connectivity of urban rail transit network, each rail station st_w is definitely affected by its adjacent rail stations $RailADST_w$, i.e., $PFlow_{i-1,j}^{RailADST_w}$. Figure 5 and Figure 6 respectively show the correlation between the current period passenger flow of *Lianban* station and the passenger flow of its adjacent two rail stations *Hubindonglu* station and *Lianhuakou* station in the previous period.

It is observed that there is a roughly linear correlation. Hence, we regard the passenger flow of the two rail stations in the previous period as two influence factors, and

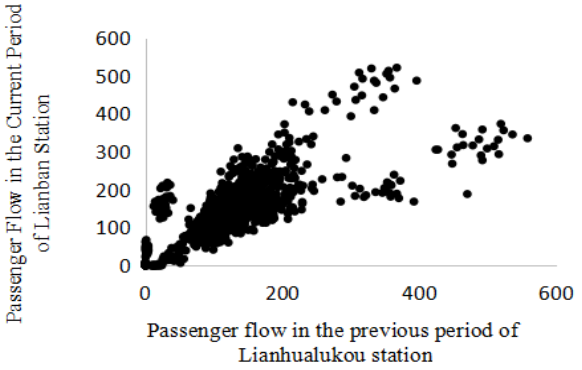

FIGURE 6. The correlation diagram between the current passenger flow of *Lianban* Station and the previous passenger flow of *Liabhualukou* rail station.

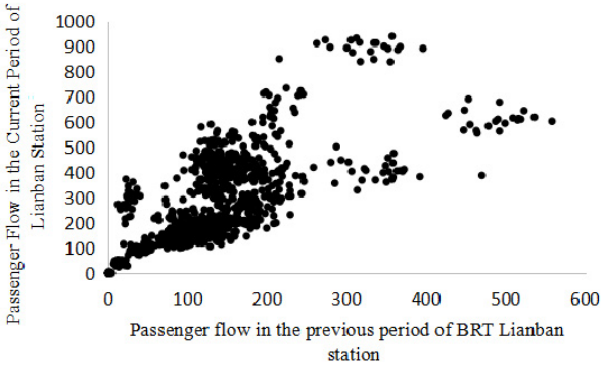

FIGURE 7. The correlation diagram between the current passenger flow of Lianban rail station and the previous passenger flow of BRT Lianban station.

use spearman method to analysis the impact of them on the passenger flow of *Lianban* station, and the result are shown in the first 2 rows of Table 3, which confirm our assumption that the passenger flow of a rail station is greatly affected by its adjacent stations.

(2) Due to that bus rapid transit (BRT) in *Xiamen* is a quite important transport, and works very similar to railway Both of BRT and rail way together constitute the main public transport network of the city, and there is an interactive relationship between them, for example, some passengers transfer from BRT to railway or from railway to BRT. Therefore, Similar to the impact of adjacent rail stations, the adjacent BRT stations $BrtADST_w$ also affects st_w, i.e., $pflow_{i,j}^{st_w}$ is affected by $Pflow_{i-1,j}^{BrtADST_w}$.

Figure 7 shows the correlation between the current period passenger flow of *Lianban* station and the previous passenger flow of BRT *Lianban* station. It addresses that the passenger flow in the current period of *Lianban* station has a strong linear relationship with the passenger flow in the previous period of BRT *Lianban* station, and the correlation coefficient is shown in the 3th row of Table 3, which addresses that BRT also has strong impact on passenger flow of rail transit.

C. FEATURE CLUSTERING

From the mentioned above, time is an important influence factor of passenger flow, and each day includes q segments, each of which has the same time interval Δt. The passenger flow increases sharply at peak, meanwhile it has little changes on the other time segments. Therefore, it is reasonable to classify time feature into a few categories $Catg_{T_i}$, which has quite different effect on passenger flow, to simplify the proposed model. Hence, k-means clustering is used to classify the time feature in this paper, as below.

Firstly, given $\Delta t = 0.5$ hour, a day is divided into 48 segments, and a railway station st_w as well as its passenger flow data within m days as a matrix:

$$PFM_{st_w} = \begin{bmatrix} pflow_{1,1}^{st_w} & pflow_{2,1}^{st_w} & \cdots & pflow_{48,1}^{st_w} \\ pflow_{1,2}^{st_w} & pflow_{2,2}^{st_w} & \cdots & pflow_{48,2}^{st_w} \\ & \cdots\cdots & & \\ pflow_{1,m}^{st_w} & pflow_{2,m}^{st_w} & \cdots & pflow_{48,m}^{st_w} \end{bmatrix} \quad (1)$$

where each row represents a passenger flow vector of a day, and each column also means a passenger flow vector composed of m days for a station at the same time segment. Here, we classify q columns into different categories.

Secondly, before clustering, we also normalized PFM first according to formula (2) to even the influence of the base value of data on the clustering effect.

$$D^* = (D - \mu)/\sigma \quad (2)$$

where D is the value before standardization, D^* is the value after standardization, μ is the mean value and σ is the variance.

Thirdly, as shown in Figure 2, there is a section rising part and descending part on both sides of the morning and evening peak, respectively. Therefore, it can be roughly classified into 4 different kinds. Hence, we set K =4 value of k-means, and run it to classify the 48 time segments, and the result is shown in Table 4 and Figure. 8.

It obviously addresses that the time segments with no passenger flow were classified into category "0" and plotted by blue curve; the time segments with lower passenger flow but changes rapidly were assigned into category "1", as the two green parts shows; the time segments with medium passenger flow and little changes were labeled as "2", and the time segments with high passenger flow were cluster "3" which are called flow peak, as the two red curves demonstrates. Hence, the category of each time segment is:

$$Catg_{T_i} = \begin{cases} 0, & i \in [0, 12) \cup (46, 47] \\ 1, & i \in [12, 15) \cup [43, 46] \\ 2, & i \in [18, 33) \cup [39, 43) \\ 3, & i \in [16, 18) \cup [33, 39) \end{cases} \quad (3)$$

D. FORECAST MODEL BASED ON LSTM NETWORK

After the data processing and feature extraction mention above, we would like to introduce multi-layer LSTM

TABLE 4. Clustering results of k-means algorithm.

Time segment	Category label	Time segment	Category label	Time segment	Category label
1	0	17	3	33	2
2	0	18	2	34	3
3	0	19	2	35	3
4	0	20	2	36	3
5	0	21	2	37	3
6	0	22	2	38	3
7	0	23	2	39	3
8	0	24	2	40	2
9	0	25	2	41	2
10	0	26	2	42	2
11	0	27	2	43	2
12	0	28	2	44	1
13	1	29	2	45	1
14	1	30	2	46	0
15	1	31	2	47	0
16	3	32	2	48	0

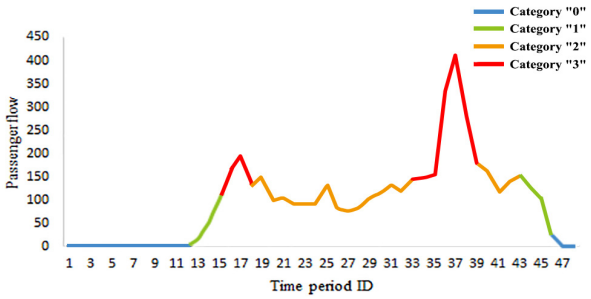

FIGURE 8. The change of entrance passenger flow in a single day at *Lianban* rail station.

Network to build a forecast model to predict passenger flow for a rail station.

1) LSTM NETWORK

Time series problems are common in daily life, such as financial market [20], industrial production, and meteorological research, etc. The recurrent neural network (RNN) model is a commonly used in deep learning model. RNN model introduces the concept of time series to solve a problem that it is difficult to understand the relationship between the input and output data, and make previous outs have a direct impact on the current input, which is quite suitable for data of long-term time-series.

However, it is difficult to train for RNN in the case of processing long time series data, and there exist problems of gradient explosion or disappearance [21]. Fortunately, Long short term memory network (LSTM), which is a variety of RNN, introduces "cell state" to update or retain historical information, so that some meaningful states can be saved for a long time. Hence, it is great for the matter of predicting passenger flow.

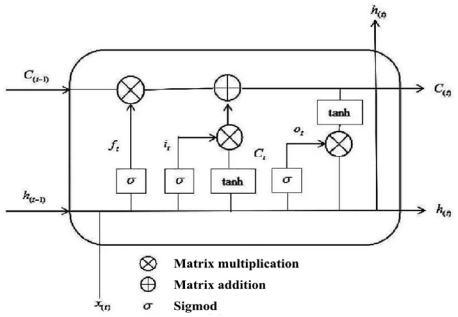

FIGURE 9. LSTM cell structure [9].

As shown in Figure 9, an LSTM unit has three gate structures, namely input gate, forgetting gate and output gate, and a cell state, which are represented by i_t, f_t, o_t, and C_t, respectively. The details of LSTM updating are shown as follows:

At first, at time t, the input gate takes the previous output h_{t-1} and the current data x_t as inputs, and outputs i_t according to formula (4), which is used to decide whether to update the information into the cell state.

$$i_t = sigmoid(W_{x\_i}x_t + W_{h\_i}h_{t-1} + b_i) \tag{4}$$

The forgetting gate takes the previous output h_{t-1} of the LSTM hidden layer and the current data x_t as the inputs. The activation function *sigmoid* makes output f_t of the forgetting gate falls within [0, 1]. If the output is 0, all previous information are discarded, otherwise, they will be retained.

$$f_t = sigmoid(W_{x\_f}x_t + W_{h\_f}h_{t-1} + b_f) \tag{5}$$

The current time candidate memory values $\tilde{C}_t$ and tanh function are used to control which new information should be added:

$$\tilde{C}_t = \tanh(W_{x\_c}x_t + W_{h\_c}h_{t-1} + b_c) \tag{6}$$

The current cell state value C_t is composed of the candidate memory value $\tilde{C}_t$ and the previous state C_{t1} according to the formula (8).

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \tag{7}$$

The output gate o_t is used to control the output of C_t. It multiplies the activation function *sigmoid* to determine which information of cell state should be output, as shown in formula (8).

$$o_t = sigmoid(W_{x\_o}x_t + W_{h\_o}h_{t-1} + b_o) \tag{8}$$

Finally, the output of the hidden layer is obtained by formula (9).

$$h_t = \mathrm{o}_t \tanh(C_t) \tag{9}$$

Among them, W_{x_i}, W_{x_f}, W_{x_o} and W_{x_c} represent the weight parameters from input layer to input gate, forgetting gate, output gate and cell state respectively. W_{h_i}, W_{h_f}, W_{h_o} and W_{h_c} represent the weight parameters from hidden layer

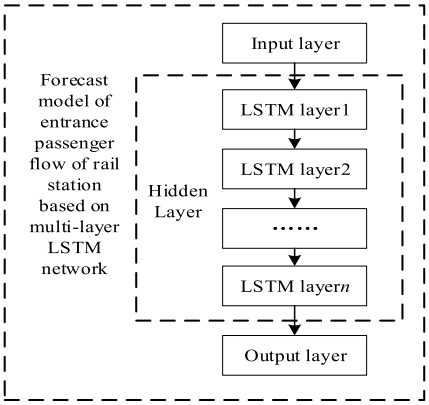

FIGURE 10. Passenger flow prediction model of rail station based on multi-layer LSTM network.

to input gate, forgetting gate, output gate and cell state respectively. b_i, b_f, b_o and b_c are bias parameters of input gate, forgetting gate, output gate and cell state respectively.

2) PASSENGER FLOW PREDICTION MODEL BASED ON MULTI-LAYER LSTM NETWORK

Although a single layer LSTM works well for many applications, we find that it obtains low accuracy for predicting passenger flow. Hence, similar to Salman *et al.* [22], we argued that multi-layer LSTM is necessary for our task.

As shown in Figure 10, the proposed multi-layer LSTM consists of input layer, hidden layer and output layer, where:

(1)The input layer is an interface that accepts parameters for station st_w:

$$Fac_{i,j}^{st_w} = \left\{ \begin{array}{l} Catg_{T_i}, Week\left(day_j\right), pflow_{i-1,j}^{st_w}, \\ pflow_{i-1,j}^{RailADST_w}, pflow_{i-1,j}^{BrtADST_w} \end{array} \right\}.$$

(2) The hidden layer in the middle is the core, which contains n LSTM layers. Each layer puts previous information forward in the form of memory stream for next step, and affects new input and output of next step.

(3) The output layer is a neuron used to predict $pflow_{i,j}^{st_w}$.

V. EXPERIMENTS

A. DATA SET AND PARAMETERS SET UP

Because *Lianban* rail station (as shown in Figure 11) is an important station in Xiamen rail line 1 that has large and stable passenger flow, it is chosen as the target station to conduct experiments.

The data sets we used in the following experiments are all normalized according to Equation (2), and the detail of are list as below.

1) **Target station**: *Lianban* rail station of rail line 1,
2) **Adjacent rail stations**: *Hubindonglu* and *Lianhualukou* station,
3) **Adjacent BRT station**: *Lianban* BRT station,
4) **Days**: July 1 2018 - July 30, 2018,
5) **Total passenger flow of 3 rail stations**: 427,027

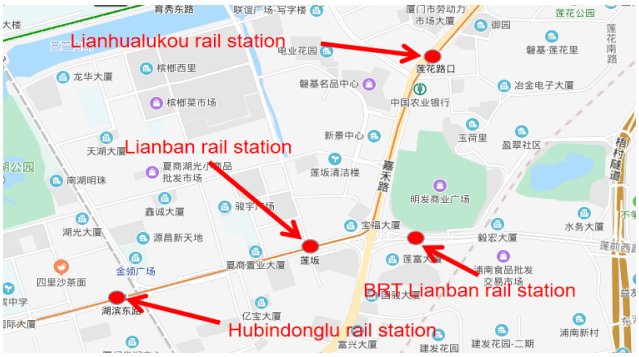

FIGURE 11. The spatial location of *Lianban* rail station.

6) **Total passenger flow of *Lianban* BRT**: 292,237
7) **Time interval**: $\Delta t = 0.5$ hours,
8) ***K* value of k-means**: $K = 4$.

In addition, the whole data set is divided into two parts, i.e., training set and test set, where training set includes the first 20 days, and the rest are testing set.

B. EVALUATION METHODS

In order to better analyze and compare the prediction effect of each experiment, based on absolute passenger error *Err* as shown in equation (10), two well-known error evaluation indexes are used, i.e. the calculation formulas of mean absolute error (*MAE*) and mean square error (*MSE*), which are shown in equation (11) and (12), respectively,

$$Err_u = \left|R_u - R_u^*\right| \tag{10}$$

$$MAE = \frac{1}{n}\sum_{u=1}^{n} Err_u \tag{11}$$

$$MSE = \frac{1}{n}\sum_{u=1}^{n} Err_u^2 \tag{12}$$

where R_u represents the actual ~~value~~ passenger flow at the u^{th} time period, R_u^* represents the predicted flow, and n represents the number of samples. The lower the values of MAE and MSE, the higher the prediction accuracy of the model.

C. THE SETUP OF MULTI-LAYER LSTM

As we know, the number of neurons of a LSTM has great impact on the result, hence, in this paper, we conduct a set of experiments to determine it for each LSTM layer.

As Table 5 shows, we can see that the Multi-Layer LSTM model with three layers, each of which has 100 neurons, obtains the best result. Hence, we will still these parameters in the following experiments.

D. COMPARISONS WITH OTHER PREDICTION MODEL

There are some other famous prediction model in this field, such as SARIMA [11], SVR [23] and BP network [24]. In this part, we make comparisons with them, as Table 6 shows. It is observed that our model based on multi-layer LSTM with

TABLE 5. Comparison of the number of neurons in each hidden layer.

NO.1 Layer neuron	NO.2 Layer neuron	NO.3 Layer neuron	NO.4 Layer neuron	*MAE*	*MSE*
50	50	50	/	12.534	386.675
100	100	100	/	11.259	302.147
150	150	150	/	12.405	319.964
100	/	/	/	13.600	473.392
100	100	/	/	12.717	398.199
100	100	50	50	14.360	403.362

TABLE 6. Comparison of the number of neurons in each layer.

Experiment	Algorithm with different setup	*MAE*	*MSE*
1	Prediction of multi-layer LSTM model + Adjacent rail stations + Adjacent BRT stations	11.259	302.147
2	prediction of using only the rail transit card swipe data	12.267	312.277
3	Prediction without k-means clustering	13.476	495.573
4	Prediction of SARIMA model[12]	15.953	887.869
5	Prediction of SVR model[23]	25.958	1573.660
6	Prediction of BP model [24]	23.198	1775.872

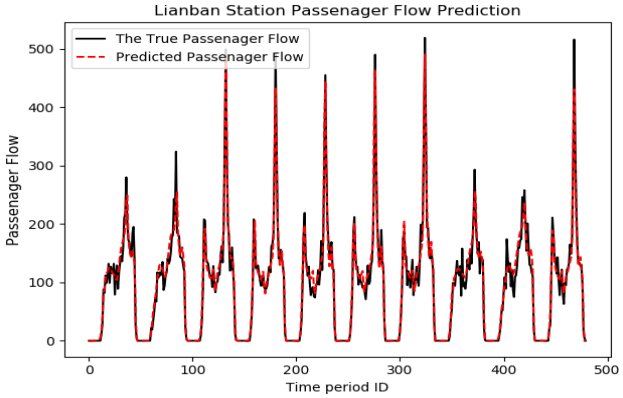

FIGURE 12. Prediction results of multi-layer LSTM model.

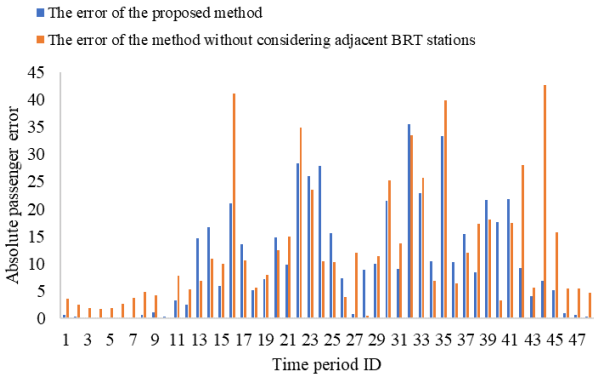

FIGURE 13. The error comparison between the proposed method and the method without considering adjacent BRT stations.

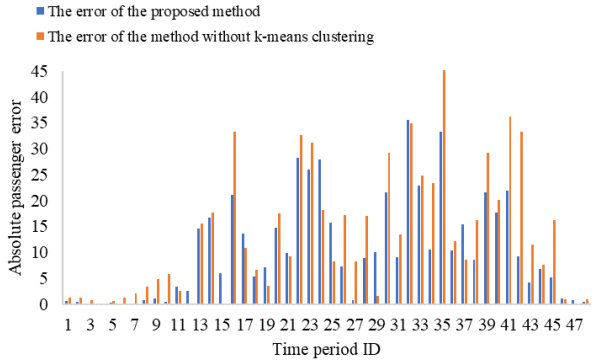

FIGURE 14. The error comparison between the proposed method and the method without k-means clustering.

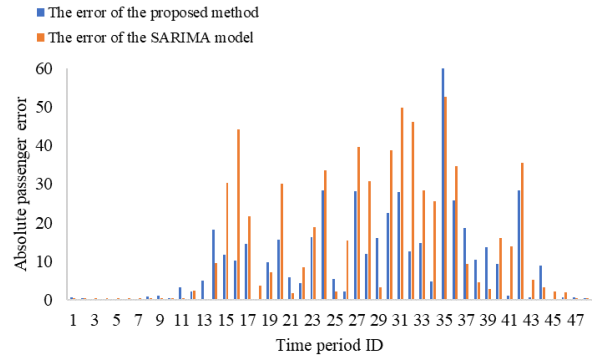

FIGURE 15. The error comparison between the proposed method and SARIMA model.

multi-source data archives the best result according to the evaluations of *MAE* and *MSE*.

The detail prediction results of each algorithm with different setup are shown as below.

(1) Figure 12 shows the results of the proposed method with multi-source data, where the black solid line is the actual data and the red dashed line is the predicted value.

(2) Figure 13 plots the passenger flow error comparisons between the proposed method and the method without considering adjacent BRT stations. It is observed that the proposed method obtains better prediction than its competitor at most time periods, which proves that adjacent BRT stations are helpful for predicting the passenger flow of the target rail station.

(3) Figure 14 shows the error comparisons between the proposed method and the method without k-means clustering, and we can see that the proposed method has more obvious superiority to the latter, which demonstrates that the classification of time periods works in the proposed algorithm.

(4) We also compare the proposed method with SARIMA, SVR and BP network model, and the results are shown in

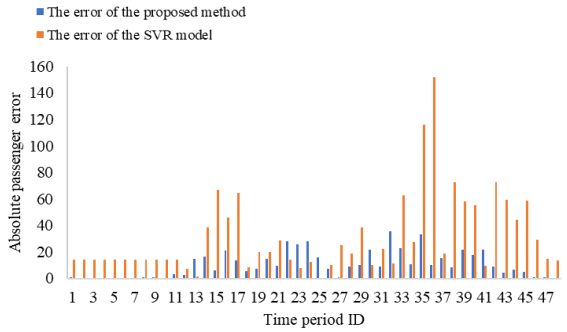

FIGURE 16. The error comparison between the proposed method and SVR model.

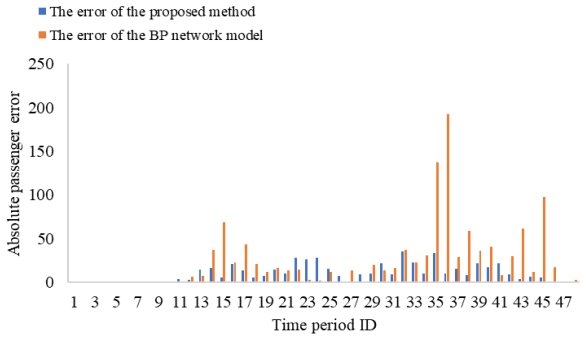

FIGURE 17. The error comparison between the proposed method and BP network model.

Figure 15, Figure 16 and Figure 17, respectively. We can witness that our method outperforms them remarkably.

VI. CONCLUSION AND FUTURE WORK

In this paper, based on multi-source data and Multi-LSTM network, a prediction method for predicting entrance passenger flow is proposed which comprehensively takes the spatial factors (including adjacent rail stations and adjacent BRT stations) and temporal factors (previous period passenger flow, weekly information and time) into considerations. Experiments shows that the proposed method is promising for predicting passenger flow, and is helpful for improving the management of city traffic control.

In the future, on the basis of the existing research, we will explore the impact of other factors to improve the performance of proposed method.

REFERENCES

[1] Y. Chen, L. Zhou, S. Pei, Z. Yu, Y. Chen, X. Liu, J. Du, and N. Xiong, "KNN-BLOCK DBSCAN: Fast clustering for large-scale data," *IEEE Trans. Syst., Man, Cybern., Syst.*, to be published.

[2] Y. Chen, S. Tang, N. Bouguila, C. Wang, J. Du, and H. Li, "A fast clustering algorithm based on pruning unnecessary distance computations in DBSCAN for high-dimensional data," *Pattern Recognit.*, vol. 83, pp. 375–387, Nov. 2018.

[3] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, Jun. 2014.

[4] Y. Chen, X. Hu, W. Fan, L. Shen, Z. Zhang, X. Liu, J. Du, H. Li, Y. Chen, and H. Li, "Fast density peak clustering for large scale data based on kNN," *Knowl.-Based Syst.*, vol. 187, Jan. 2020, Art. no. 104824.

[5] Y. Liu, Z. Liu, and R. Jia, "DeepPF: A deep learning based architecture for metro passenger flow prediction," *Transp. Res. C, Emerg. Technol.*, vol. 101, pp. 18–34, Apr. 2019.

[6] Y. Jia, P. He, S. Liu, and L. Cao, "A combined forecasting model for passenger flow based on GM and ARMA," *Int. J. Hybrid Inf. Technol.*, vol. 9, no. 2, pp. 215–226, Feb. 2016.

[7] Y. Wei and M.-C. Chen, "Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks," *Transp. Res. C, Emerg. Technol.*, vol. 21, no. 1, pp. 148–162, Apr. 2012.

[8] L. Liu and R.-C. Chen, "A novel passenger flow prediction model using deep learning methods," *Transp. Res. C, Emerg. Technol.*, vol. 84, pp. 74–91, Nov. 2017.

[9] Q. Tang, M. Yang, and Y. Yang, "ST-LSTM: A deep learning approach combined spatio-temporal features for short-term forecast in rail transit," *J. Adv. Transp.*, vol. 2019, pp. 1–8, Feb. 2019.

[10] Z. Junwen, *Research and Implementation of Spatiotemporal Passenger Flow Model*. Beijing, China: Beijing Univ. of Posts and Telecommunications (in Chinese), 2018.

[11] L. Mei, L. Jing, W. Zijian, W. Sida, and C. Laijin, "Short-time passenger flow forecasting at subway station based on deep learning lstm structure," (in Chinese), *Urban Mass Transit*, vol. 21, no. 11, pp. 42–46, and 77, 2018.

[12] M. Milenković, L. Švadlenka, V. Melichar, N. Bojović, and Z. Avramović, "SARIMA modelling approach for railway passenger flow forecasting," *Transport*, vol. 33, no. 5, pp. 1113–1120, Mar. 2016.

[13] Y. Sun, B. Leng, and W. Guan, "A novel wavelet-SVM short-time passenger flow prediction in Beijing subway system," *Neurocomputing*, vol. 166, pp. 109–121, Oct. 2015.

[14] S.-Z. Zhao, T.-H. Ni, Y. Wang, and X.-T. Gao, "A new approach to the prediction of passenger flow in a transit system," *Comput. Math. Appl.*, vol. 61, no. 8, pp. 1968–1974, Apr. 2011.

[15] X. Jiang, L. Zhang, and X. Chen, "Short-term forecasting of high-speed rail demand: A hybrid approach combining ensemble empirical mode decomposition and gray support vector machine with real-world applications in China," *Transp. Res. C, Emerg. Technol.*, vol. 44, pp. 110–127, Jul. 2014.

[16] L. Xiaoqiang, L. Jie, and C. Yanru, "Metro short-term traffic flow prediction with deep learning," (in Chinese), *Control Decis.*, vol. 34, no. 8, pp. 1589–1600, 2019.

[17] P. Jiao, R. Li, T. Sun, Z. Hou, and A. Ibrahim, "Three revised Kalman filtering models for short-term rail transit passenger flow prediction," *Math. Problems Eng.*, vol. 2016, pp. 1–10, 2016.

[18] X. Wang, K. An, L. Tang, and X. Chen, "Short term prediction of freeway exiting volume based on SVM and KNN," *Int. J. Transp. Sci. Technol.*, vol. 4, no. 3, pp. 337–352, 2015.

[19] R. Chrobok, J. Wahle, and M. Schreckenberg, "Traffic forecast using simulations of large scale networks," in *Proc. IEEE Intell. Transp. Syst. (ITSC)*, Nov. 2001, pp. 434–439.

[20] S. Pei, T. Shen, X. Wang, C. Gu, Z. Ning, X. Ye, and N. Xiong, "3DACN: 3D augmented convolutional network for time series data," *Inf. Sci.*, vol. 513, pp. 17–29, Mar. 2020.

[21] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1310–1318.

[22] A. G. Salman, Y. Heryadi, E. Abdurahman, and W. Suparta, "Single layer & multi-layer long short-term memory (LSTM) model with intermediate variables for weather forecasting," *Procedia Comput. Sci.*, vol. 135, pp. 89–98, 2018.

[23] Y. Ke-wu, "Study on the forecast of air passenger flow based on SVM regression algorithm," in *Proc. 1st Int. Workshop Database Technol. Appl.*, Apr. 2009, pp. 325–328.

[24] Y. Wang, D. Zheng, S. M. Luo, D. M. Zhan, and P. Nie, "The research of railway passenger flow prediction model based on BP neural network," *Adv. Mater. Res.*, vols. 605–607, pp. 2366–2369, Dec. 2012.