A Study on Short-Term Passenger Flow Forecasting for Urban Rail Transit

Zihan Fan[1, *], Jiaqi Si[2], Yiqian Wang[3]

[1]Road Bridge and River Crossing Engineering Program of Chang'an University, Xi'an, China

[2]Information Management and Information Systems Program of Lanzhou University of Technology, Lanzhou, China

[3]School of Southern University of Science and Technology, Shenzhen, China

*Corresponding author: 2022903894@chd.edu.cn

Abstract. Short-term forecasts of light rail traffic are an indispensable basis for adjusting timetables in real time, improving operational efficiency and providing passengers with informed travel recommendations. At the same time, the subway is the most convenient and widely used mode of transport for people today. The purpose of this paper is to forecast passenger traffic in the subway. To solve the problem of predicting short-term passenger flow at the entrances and exits of metro stations with nonlinear and stochastic characteristics, this paper uses a genetic algorithm (GA) based on ordered long-term and short-term memory (SLSTM) and builds a GA-SLSTM forecasting model to predict future short-term passenger flow by studying the patterns and characteristics of passenger flow in historical data. The model consists of five stages: data preprocessing, model training, model evaluation, and forecast application. Through these steps, the GA-SLSTM model is able to predict the future situation of short-term passenger traffic based on the characteristics and patterns in historical data, providing a reference base for planning and optimizing urban transport.

Keywords: Forecasting, urban rail transit, short-term passenger traffic.

1. Introduction

In recent years, people have chosen urban rail transport for reasons of comfort, convenience, environmental friendliness, and low CO2 emissions, but high passenger traffic also brings a number of problems such as station crowding and delays during peak periods [1]. With increasing passenger traffic, short-term forecasting has become a key factor in ensuring safety and optimising operational efficiency, while providing a benchmark for planning passenger journeys [2-3]. Short-term forecasting concerns passengers with a forecasting interval of less than 15 minutes. Compared with medium- and long-term forecasting, short-term passenger traffic forecasting exhibits stochastic properties and higher volatility and the accuracy of forecasting is directly related to the efficiency of urban rail transport organisation and operation. Some researchers have investigated forecasting methods for short-term urban railway traffic forecasting using nonlinear forecasting or combinatorial optimisation models, which have demonstrated the feasibility of short-term passenger flow forecasting.

The GA-SLSTM forecasting model collects metro operation data in 10-minute forecasting steps, analyses the characteristics of passenger flow variation, and compares the forecasting effect with the GA-Recursive Neural Network (RNN) model and the LSTM model; the average coefficient of determination (R^2) value of the forecasting value of the GA-SLSTM forecasting model reaches 0.95 and 0.95 for regular stations and junction stations. The average values of the coefficient of determination (R2) of the GA-SLSTM forecasting model for regular stations and interchange stations reached 0.95 and 0.90, respectively, indicating that the forecasting value corresponds to the actual value and the forecasting error is smaller than the other two models, which indicates that this method can improve the accuracy of short-term forecasting of metro passenger traffic [4-6].

Convolutional or recurrent artificial neural network models, which are commonly used in passenger traffic prediction tasks, face the dilemma of vanishing or exploding gradients and have difficulty handling data with significant seasonality and large duration, while long- and short-term

memory models can solve these problems. Wang et al. combine the LSTM model with the adaptive K-means algorithm and show that this method is effective for passenger traffic forecasting [7]. Liu et al. improved the forecasting accuracy by dividing the historical data into current, short-term, and long-term and fed them separately to the LSTM model [8]. The choice of hyperparameters for the forecasting model is very important and affects the quality of the model, so some researchers use optimization algorithms to optimize the model. Qiu et al. optimize the kernel function and kernel parameters of support vector machine using GA [9]. Liu et al. optimize an artificial back propagation (BP) neural network using the particle swarm algorithm [10].

The GA-SLSTM model enables to improve the accuracy of short-term passenger traffic forecasting in urban rail transport. Using GA to optimize the structure of the artificial neural network makes the model more robust compared to manual parameter tuning and other methods. The optimized SLSTM prediction model is able to assimilate long sequential information and accurately predict changes in passenger flow, which is much better than other models at both ordinary and junction stations. In order to improve the accuracy of short-term passenger flow forecasting, the next step is to analyze and investigate the trends, periods and noise decomposition of passenger flow data at interchange stations to further improve the forecasting accuracy of interchange stations.

2. Methods

2.1. Data Source and Description

The research in this paper selects Beijing subway inbound and outbound data, covering May 12 to May 22, 2018, involving 47 subway lines, 646 stations, and about 60 million pieces of data. The raw data includes four types of information: time, inbound data, outbound data, and in/outbound data.

Data preprocessing is an important part before modeling. For the algorithm to be effective, it must be provided with clean, accurate and concise data. However, data collected in real-world applications usually carries errors. If missing values are not handled correctly, one may end up with inaccurate data inferences. At the same time, large amounts of redundant data increase memory consumption, and correctly processing data can reduce time, lower cost and improve model quality.

We found that the subway inbound and outbound data is basically 0 in the 0:00 to 5:00 time period, this is due to the fact that the daily maintenance time of Beijing subway stations is from 0:00 to 5:00, the inbound and outbound data during this period is generated by the staff, which has no practical significance in the subway short-term traffic flow prediction problem, as a consequence, in the data preprocessing, we discarded the invalid data generated in the 0:00 to 5:00 time period.

2.2. Indicator Selection and Description

The evaluation index equalization coefficient EC was used as the population fitness A_{EC}

$$A_{EC} = 1 - \frac{\sqrt{\sum_{i=1}^{n}(y_i' - y_i)^2}}{\sqrt{\sum_{i=1}^{n} y_i^2} + \sqrt{\sum_{i=1}^{n} y_i'}} \tag{1}$$

Where y_i is the true value and y_i' is the predicted value. $A_{EC} \in [0,1]$. The closer the value of the equality coefficient A_{EC} is to 1, the closer the predicted value is to the real value, and the higher the probability that the individual will be selected, the next generation of the population will select the individual according to the fitness and carry out the crossover and mutation operations. In this paper, we set the maximum population iteration to 20 generations, repeat the selection, crossover and mutation to generate a random population, and build the corresponding model until the maximum population iteration, and output the optimal individual with the largest A_{EC} value and the corresponding prediction model.

2.3. Research Method

Genetic algorithm is an optimisation algorithm that mimics the process of genetic evolution in nature and is designed to solve complex problems. Long-term short-term memory network is a type of recurrent neural network capable of reflecting the long-term dependence of time series data. The combination of GA algorithm and LSTM algorithm in short-term forecasting of railway passenger traffic has the following advantages: possibility of parameter optimisation, robustness of the model, non-linear modelling, modelling of long-term dependencies, solving complex problems, etc. The GA algorithm is better at solving these complex problems by re-selecting the parameter space to improve the accuracy and robustness of the prediction.

Thus, the combination of the GA algorithm and the LSTM algorithm can be useful for short-term forecasting of passenger rail traffic, improving the accuracy and robustness of forecasting, and adapting to the requirements of solving complex problems.

3. Results and Discussion

Based on the daily passenger traffic of Beijing West Railway Station, the total passenger traffic of the station from 1 May to 10 May was calculated and systematically analysed, which allowed the passenger traffic from 11 May to 20 May to be forecast and the passenger traffic changes to be evaluated and analysed in a targeted manner. Thus, the benefits of the model were demonstrated.

3.1. Problem Analysis

GA (Genetic Algorithm) is an optimization algorithm that mimics the process of genetic evolution in nature and is suitable for solving complex problems. The GA algorithm can find the optimal configuration of the LSTM model by optimizing the parameters and structure of the LSTM model, and the crossover and mutation operations in the GA search process can effectively improve the performance of the model. model long-term dependencies. This is especially important when considering the characteristics of passenger traffic over time, which can facilitate the prediction of future passenger traffic trends. The combination of the GA algorithm and the LSTM algorithm can play a useful role in short-term passenger traffic forecasting, improve the accuracy and robustness of the forecast, and adapt to the needs of solving complex problems.

3.2. LSTM Modeling

The architecture of our combined GA-LSTM model is depicted in Fig. 1

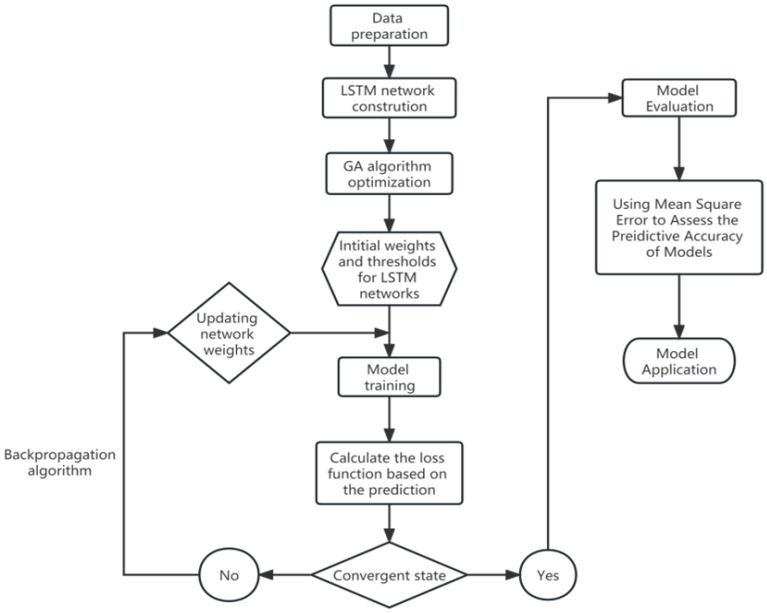


Fig. 1 Model Construction Diagram

3.3. Parametric Analysis

3.3.1 GA parameterization

The main indicators of the GA mathematical model analysis are population_size, max_generations, elite_size, and mutation_rate, as Table 1 shows.

Table 1. About parameter setting

Number	Parameter name	Reference point
1	population_size	50
2	max_generations	100
3	elite_size	10
4	mutation_rate	0.01

3.3.2 LSTM parameterization

The LSTM model has several parameterizations. The parameter "hidden_size" determines the number of LSTM neurons, which affects the model's expressive power, complexity, and training time. Increasing the hidden layer dimension improves expressiveness but increases complexity and training time. The parameter "num_layers" determines the number of layers in the LSTM network. Increasing network depth improves expressiveness, but also increases complexity and training time. The parameter "bidirectional" indicates whether to use bidirectional LSTM. Bidirectional LSTM processes both forward and reverse input sequences, capturing contextual information effectively. However, it also incurs higher computational overhead. The parameter "sequence_length" specifies the length of the input sequence. The input sequence length must correspond to the training data. Otherwise, preprocessing such as truncation or padding is needed. The parameter "input_size" specifies the dimension of the input features. The input feature dimension should match the training data, requiring preprocessing if necessary. The parameter "learning_rate" controls the update step of model parameters. The choice of learning rate significantly impacts the training process and model performance. The parameter "batch_size" determines the number of samples used in each training iteration. Batch training improves efficiency and stability but increases memory usage. The parameter "num_epochs" specifies the number of training iterations for the model. More training iterations allow the model to learn dataset features more comprehensively, but may increase the risk of overfitting.

3.4. Analysis of Results

Visualization and analysis of the predicted and true value cases resulted in the following results, as Figure 2 to 6 showing.

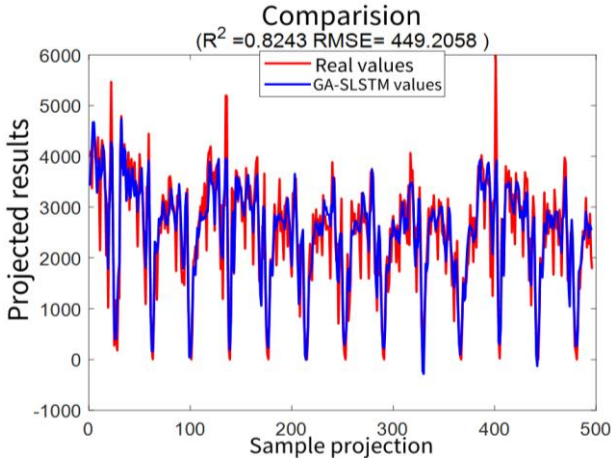


Fig. 2 Training Set Prediction Graph

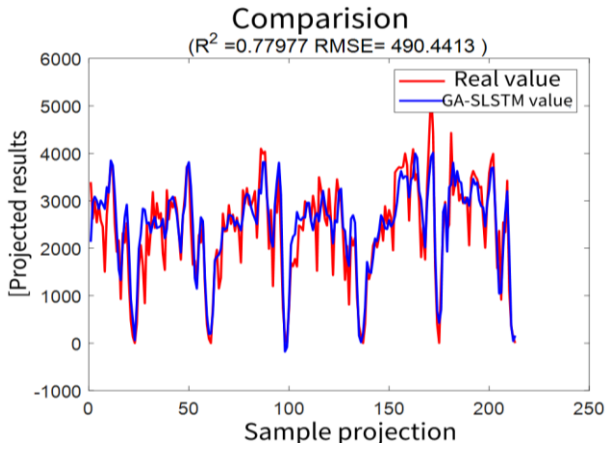


Fig. 3 Test Set Prediction Graph

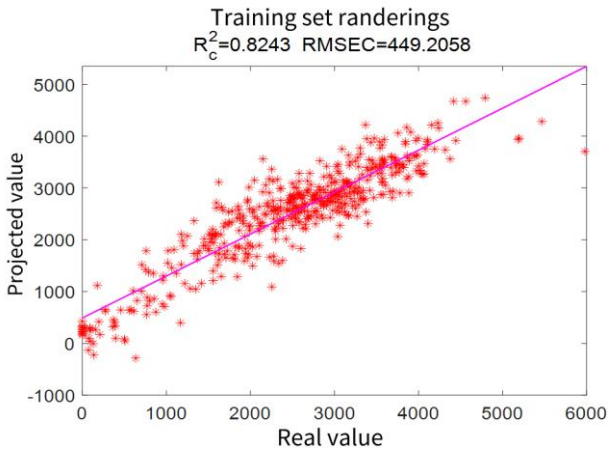


Fig. 4 Test Set Prediction Graph

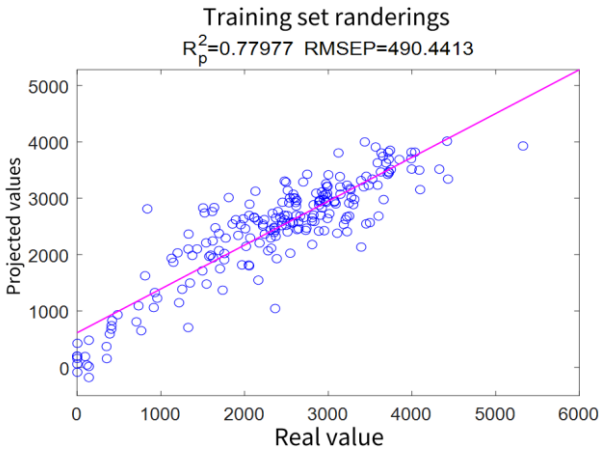


Fig. 5 Test Set Effect Image

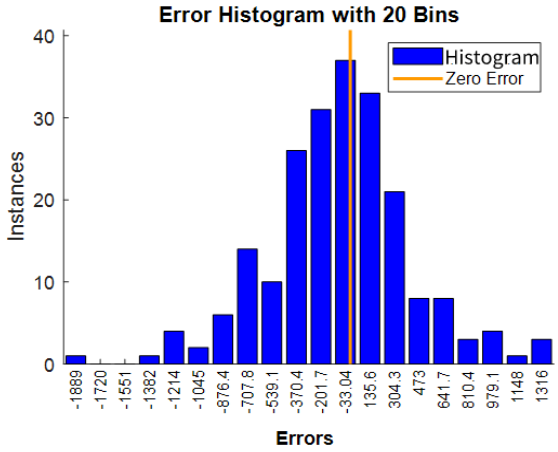


Fig. 6 Error histogram

From the figure 2, 3, 4, 5 and 6, it can be seen that for a given metro passenger flow prediction site, GA-LSTM is able to predict the cyclical characteristics of metro passenger flow changes and peak period characteristics to a certain extent, and is able to accurately predict the peak concentration time of the metro passenger flow, and the predicted value of the model is more consistent with the actual value, but the peak effect peak dispersion value is poor, and it cannot accurately predict the curve of the change of passenger flow. At the same time, the prediction results of the LSTM model were analysed using the evaluation criteria. The results are summarised in Table 2.

Table 2. GA-LSTM model prediction results

Number	Type	value
1	MAE	385
2	RPD	2.0747
3	R2	0.76535

3.5. GA-LSTM Model Based Prediction

3.5.1 Problem analysis

Only short-time passenger flow prediction is not intuitive enough to guide pedestrians on the subway, with the development of intelligent transportation, it is inevitable that the subway short-time congestion prediction puts forward higher requirements. On the basis of short-time passenger flow prediction, we have set up evaluation indexes and programmed, in order to more accurately and quickly derive the short-time congestion situation of subway stations.

3.5.2 Indicator setting

We set 0 to 1,000 passengers in a half hour as smooth, 1,000 to 2,000 passengers in a half hour as normal, 2,000 to 3,000 passengers in a half hour as slightly congested, 3,000 to 5,000 passengers in a half hour as congested, and 5,000 and more passengers in a half hour as very congested.

3.5.3 Example and analysis of forecast results

It is shown that our GA-LSTM model can effectively predict the short-time subway congestion, and Table 3 shows some of the results.

Table 3. Congestion forecast tables

Time	Predicting congestion levels	Actual congestion level
12:30	congested	congested
13:00	congested	congested
13:30	congested	congested
14:00	congested	slightly congested
14:30	congested	congested
15:00:	congested	congested
15:30	slightly congested	congested
16:00	slightly congested	slightly congested
16:30	slightly congested	slightly congested
17:00	slightly congested	congested
17:30	congested	congested
18:00	congested	congested
18:30	slightly congested	congested
19:00	normal	normal

4. Conclusion

This article summarizes the features and benefits of the GA-SLSTM with the aim of being a reference and guide for its practical application. First, the basic principles and background of the GA-SLSTM are introduced. The GA-SLSTM is a time-series-based model for short-term forecasts of future passenger flows. The model takes into account the effect of holidays, weather conditions, geographical location and other factors on passenger numbers. It is suitable for forecasting passenger flows in the metro. Secondly, this article selected data from Beijing metro stations, determined the indicators and analyzed the benefits of the model. Finally, we conducted the corresponding analysis and used the model to analyze the parameters and predict passenger flow. Although the model has high accuracy and reliability, there are still some limitations in its practical application, such as the long training time of the model and the high computer usage. This paper wants to improve the model as much as possible to give users the best experience. In the future, this paper wants to investigate more relevant stations, build a short-term congestion prediction network and combine it with travel content to offer pedestrian services in the form of apps for optimal metro station selection.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Zhong C, Batty M, Manley E, et al. Variability in Regularity: Mining Temporal Mobility Patterns in London, Singapore and Beijing Using Smart-Card Data. PLoS ONE, 2016, 11(2): e0149222.

[2] Niu H, Zhou X. Urban railway timetable optimisation under time-varying demand and congestion.Transportation Research Part C: Emerging Technologies, 2013, (36): 212-230.

[3] Wu J, Liu M, Sun H, et al. Fairness-based schedule synchronisation optimisation in urban metro networks. Transport research part C: new technologies, 2015, (51): 1-18.

[4] TENG Teng, LIU Zhengqi, WANG Xiaomin. Short-term forecasting of urban railway passenger traffic based on GA-SLSTM model. Railway Computer Software, 2022, 31(08): 7-12.

[5] He Xudong. Analysis of urban railway passenger traffic forecasting methods using big data. Science and Technology Information, 2021, 19(30): 11-13.

[6] QIN Linan, DONG Luxi. Research on urban railway passenger traffic forecasting algorithm. Transport Engineering, 2021, 21(01): 40-47.

[7] Keyu Wen, Guotang Zhao, Bisheng He, et al. An information transfer decomposition-based forecasting method for short-distance railway holiday passenger traffic. Expert Systems with Applications, 2022, 189 (3).

[8] Chen S J, Xue Y. Passenger flow prediction for short-distance bus journeys based on an improved convolutional neural network. Computer Science, 2019, 46: 175-184.

[9] Wang Qiuwen, Chen Yanru, Liu Yuanchun. Short-term passenger flow prediction of urban railway travel based on a convolutional neural network with long and short-term memory. Control and Decision Making, 2021, 36: 2760-2770.

[10] Liu Yu et al. Deep learning-based architecture for underground passenger flow prediction. Transportation Research Part C: Emerging Technologies, 2019, (101): 18-34.