Short-term Passenger Flow Forecast of Urban Subway Transportation Based on Deep Learning Methods

Yueyan Lyu1*

1Southeast University, China,211100

*Corresponding author e-mail: 213180955@seu.edu.cn

Abstract. With the acceleration of urban construction, many urban rail transits has entered the stage of network development. Facing the rail passenger flow with larger flow and more complex characteristics, how to forecast the rail passenger flow in real time and accurately has become an important topic in the field of transportation planning. Rail passenger flow is closely related to the travel patterns of urban residents, and is affected by factors such as weather conditions and station development. Based on the heterogeneous data such as rail passenger flow data, POI data, meteorological data, air quality data and road network distribution data, this article analyzes the spatial and temporal distribution characteristics of rail passenger flow, conducts functional clustering analysis of stations, and constructs hybrid neural network using deep learning method to realize high-precision prediction of short-term passenger flow at rail stations.

Keywords: Short-term Passenger Flow Forecast, Subway Transportation, Deep Learning

1. AFC data and processing

This study is based on the AFC data of Suzhou Rail Transit Line 1, 2 and 4, the data mainly includes the post-desensitization ID, the date, time and site of the entry and the date, time and site of the exit.

After completing the cleaning of the abnormal AFC data, taking 15 minutes as the time granularity and taking the time granularity as one of the parameters in the function, output the passenger flow in and out of each station, and each record value contains the ID of all sites and outputs it for spatial and temporal characteristics analysis[1]. After analysis, it is found that the distribution law of passenger flow of urban rail transit is significantly different: On weekdays, when passengers travel, there will be peaks in the morning and evening, and there are rules to follow; On weekends and holidays, the fluctuation of passenger flow is more complicated, which is caused by the complicated travel purpose.
2. Site clustering

The classification of urban rail transit stations plays an important role in the study of land use, passenger flow change and development trend around different types of stations. The site clustering method can be analyzed from the perspective of the land use around the site.

Cluster analysis is an unsupervised machine learning method, whose core is the process of dividing the data set into several subsets, so that the data within the same set has a high similarity, while the data between different sets have obvious differences[2]. The site function clustering methods adopted in this study include k-means, mean-shift and spectral clustering.

2.1. Selection and analysis of clustering methods

Based on POI data, this study uses a method based on the percentage of points of interest to cluster the sites, and characterize the site functions according to the clustering results.

The percentage of interest points can depict the functional characteristics of the site more intuitively, which can describe the distribution of different types of interest points around the site, which is directly related to the land use division around the site.

\[
p_{i,k} = \frac{N_{i,k}}{N_k}
\]

(2-1)

By processing the data, this study can get the percentage of POI in each category of each site, then k-means, mean-shift and spectral clustering are used for clustering analysis.

Cluster analysis is carried out on 92 sites and 12 points of interest.

2.2. k-means clustering

The basic idea of k-means algorithm is to divide the data into predetermined class number K through continuous iteration on the basis of minimizing the error function[3]. The k-means algorithm generally adopts the sum of error squares as the standard measurement function, which is defined as follows:

\[
SSE = \sum_{i=1}^{K} \sum_{x \in B_i} |x - \bar{x}_i|^2
\]

(2-2)

In the formula, SSE represents the sum of error squares between all objects in the set and the center of their subset; \(x\) is a point in the object space; \(\bar{x}_i\) is the mean value of the cluster. The similarity measure of the k-means algorithm uses the Euclidean distance method, and its calculation formula is:

\[
d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}
\]

(2-3)

In the formula, \(d(x,y)\) represents the distance between \(x(x_1,x_2,\cdots,x_n)\) and \(y(y_1,y_2,\cdots,y_n)\) in an n-dimensional object.

2.3. Mean-shift clustering

The core idea of mean-shift algorithm is to constantly find new coordinates of the center of the circle until the region with the highest density[4].
Estimate the window radius using sklearn’s `estimate_bandwidth`, the relation between the value of parameter `quantile` (bandwidth value quantile) and the number of result categories is as follows:

![Figure 2.1. Mean-shift clustering results](image)

When `quantile` reaches below 0.2, the clustering number reaches a more reasonable value. It's been tested and found that if `quantile` continues to decrease, the number of categories will continue to increase, remaining at the longest range of 9. Therefore, it can take `quantile`=0.18, and the clustering can be 9 classes.

2.4. **Spectral clustering and comparison**

The basic idea of Spectral clustering is to use the similarity matrix of sample data for eigende composition and cluster the obtained eigenvectors, which has nothing to do with the characteristics of samples but only with the number of samples[5].

In spectral clustering, the clustering number is 7, 8 and 9, respectively, and compare the results with the results of k-means taking 7 and 8 categories and mean shift taking 9 categories. It can be found that when the clustering number is 7, the clustering results are relatively consistent.

2.5. **Comparison of clustering results**

Based on the results of the three clustering methods, the comparison is made as follows:

![Figure 2.2. Clustering results: k-means(left) and mean-shift(right)](image)
The passenger flow forecast results and discussion

3.1. Passenger flow forecast

In the early stage of this study, a multi-input-single output passenger flow prediction model based on multi-source data is established. Firstly, the observed value of orbital passenger flow is converted into the passenger flow input of the model, the multi-layer one-dimensional convolutional network and cyclic network are used to extract the spatio-temporal evolution characteristics of passenger flow. Then, the multilayer perceptron is used to process the external environmental factors and merge them into the fusion layer together with the cyclic layer output to measure the predicted value of the orbital passenger flow, the average absolute error is used to measure the prediction effect of the model.
3.2. **Discussion on passenger flow forecast results**

The following shows the comparison between the predicted and observed values of orbital passenger flow, it can be found that the model can fit the variation law of passenger flow well, and the prediction accuracy is high. Based on the results of site clustering, Figure 3.4 shows that first classify the sites, and then build models for each category of sites, compared with the direct prediction without classification, the result is significantly improved.

**Figure 3.1.** Deep learning model for multi-input-single output passenger flow prediction

**Figure 3.2.** Comparison chart of predicted passenger flow and observed passenger flow from September 20th to September 30th at Jinfeng Road Station
4. Conclusion

Based on the in-depth analysis of the AFC data of Suzhou urban rail transit, this study establishes the LSTM urban rail transit short-time passenger flow prediction model, the model is verified by taking the inbound volume prediction of Jinfeng Road Station of Line 1 and Likou Station of Line 2 as an example, the model can fit the variation law of the passenger flow of the station well, and the prediction accuracy is high.

References

[1] Voort M V D, Dougherty M, Watson S. Combining kohonen maps with arima tme series models to forecast traffic flow [J]. Transportation Research Part C merging Technologies, 1996, 4(5): 307-318.

[2] Kolen J, Kremer S. Gradient Flow in Recurrent Nets: The Difficulty of Learning Long Term Dependencies [M]. Wiley-IEEE Press, 2001.
[3] Ma X, Tao Z, Wang Y, et al. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data [J]. Transportation Research Part C, 2015, 54: 187-197.

[4] P. Poonia and V. K. Jain, "Short-Term Traffic Flow Prediction: Using LSTM," 2020 International Conference on Emerging Trends in Communication, Control and Computing (ICONC3), Lakshmangarh, Sikar, India, 2020, pp. 1-4, doi: 10.1109/ICONC345789.2020.9117329.

[5] R. Fu, Z. Zhang and L. Li, "Using LSTM and GRU neural network methods for traffic flow prediction," 2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC), Wuhan, 2016, pp. 324-328, doi: 10.1109/YAC.2016.7804912.