Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 1055910, 14 pages
https://doi.org/10.1155/2021/1055910

Research Article

Passenger Flow Prediction of Integrated Passenger Terminal Based on *K*-Means–GRNN

Yifan Tan,[1,2,3] **Haixu Liu**,[1,2,3] **Yun Pu**,[1,2,3] **Xuemei Wu**,[1,2,3] **and Yubo Jiao**[1,2,3]

[1]*School of Transportation and Logistics SWJTU, Chengdu 610031, China*
[2]*National United Engineering Laboratory of Integrated and Intelligent Transportation SWJTU, Chengdu 610031, China*
[3]*National Engineering Laboratory of Application Technology of Integrated Transportation Big Data SWJTU, Chengdu 610031, China*

Correspondence should be addressed to Haixu Liu; hxliu@swjtu.edu.cn

Received 17 April 2021; Revised 23 July 2021; Accepted 28 August 2021; Published 8 October 2021

Academic Editor: Arkatkar Shriniwas

As the passenger flow distribution center cooperating with various modes of transportation, the comprehensive passenger transport hub brings convenience to passengers. With the diversification of passenger travel modes, the passenger flow scale gradually increases, which brings significant challenges to the integrated passenger hub. Therefore, it is urgent to solve the problem of early warning and response to the future passenger flow to avoid congestion accidents. In this paper, the passenger flow GRNN prediction model is proposed, based on the *K*-means cluster algorithm, and an improved index named BWPs (Between-Within Proportion-Similarity) is proposed to improve the clustering effect of *K*-means so that the clustering effect of the new index is verified. In addition, the passenger flow data are studied and trained by combining with the GRNN neural network model based on parameter optimization (GA); the passenger flow prediction model is obtained. Finally, the passenger flow of Chengdu East Railway Station has been taken as an example, which is divided into 16 models, and each type of passenger flow is predicted, respectively. Compared with the traditional method, the results show that the model can predict the passenger flow with high accuracy.

1. Introduction

As an important form of the urban intelligent transportation system [1], the comprehensive passenger transport hub carries the passenger flow of various transportation modes and connects to an urban transportation network that carries out integration and conversion [2]. It is a vertical facility that realizes zero-distance transfer and plays an important role in urban transportation and urban long-term construction planning.

With the rapid growth of passenger traffic demand, the carrying capacity is gradually challenged, and the internal passenger flow types are characterized by high complexity, strong randomness, significant cross conflicts, and many hidden nodes [3].

To relieve the pressure of regional transportation and avoid the formation of high-density passenger flow gathering areas, it is necessary to give an early warning based on the passenger flow quickly forecast. Effective prediction can improve the proportion of public transport and promote the coordinated development of urban culture, economy, and environment. So it is urgent to forecast the large passenger flow in the comprehensive passenger transport hub in advance, which can effectively increase the proportion of public transportation and promote the coordinated development of urban culture, economy, and environment.

Scholars often make a joint prediction of short-term passenger flow with the help of other relevant real-time data, such as the changing weather, holidays, events, surrounding traffic conditions, and other factors. Multiple modal data prediction models often need data support from multiple platforms. Although the accuracy of prediction is improved, the prediction efficiency is low, which tends to ignore the timeliness of short-term prediction. The multimodal prediction model is more suitable for providing additional suggestions for urban transit hub planning and construction. Moreover, multimodal data prediction requires multiple platforms, which will cause problems such as increased operating costs and long prediction time. At the same time, the data collected by the existing integrated passenger terminal platform is extremely limited, mainly including the time-series AFC card flow data and real-time video image data, which are not connected to other transportation facilities platforms and cannot be combined for multiplatform prediction.

At the same time, the passenger flow warning is mainly carried out through video monitoring. The real-time monitoring of the passenger flow area is carried out to identify and process the passenger flow state [4]. Therefore, it is hard to achieve effective prejudgment to take reasonable measures, which is an urgent problem for us to make a rapid and accurate prediction based on the time-series data.

It is important to classify the passenger flow of the comprehensive passenger terminal and then identify complex passenger flow types to construct the prediction model for different passenger flow types, which makes real-time prediction analysis. It provides a new idea and new method for predicting the passenger flow state of the comprehensive passenger terminal in the future.

The main contributions of this paper are as follows:

(1) To classify complex passenger flows, BWPs index is proposed to enhance the clustering effect of the *K*-means algorithm.

(2) Verification of the clustering performance of the BWPs index is done, which is compared with the classical clustering index.

(3) The model of GRNN is proposed to predict the daily passenger flow, which is based on *K*-means, and its structure has been defined.

(4) Demonstration of the effectiveness of the model and the new cluster index is done using real-life case studies.

(5) The passenger flow of Chengdu East Railway Station has been divided into 16 types, and the trend of passenger flow was analyzed according to passenger flow labels.

The remaining parts of the paper are organized as follows. Section 2 briefly reviews the existing literature. BWPs index is proposed, which has been verified based on the *K*-means clustering algorithm in Section 3. Section 4 describes the *K*-means–GRNN algorithm. Section 5 is the algorithm example analysis. Finally, conclusions are summarized in Section 6.

2. Literature Review

In recent years, it is a common method of forecasting by correlation analysis and similarity analysis, such as the regression model [5], ARIMA model [6, 7], and MLE (Maximum Likelihood Estimation) [8], which is statistics-based models. These methods rely on historical data and cannot solve the problems of randomness, time-variability, and uncertainty of passenger flow, which makes it hard to express some nonlinear characteristics of data. The timeliness cannot be satisfied, neither is real-time online forecasting. Some studies have proposed nonlinear prediction models, such as supporting vector machines (SVM) [9] and neural networks [10], which are machine-learning-based models. Recently, deep-learning-based models have been widely introduced to tackle this problem and have been proved to have great advantages over previous models.

For example, the long short-term memory (LSTM) network [11], which is based on RNN [12], was introduced to predict traffic speed at first, and some scholar has enhanced long-term features to improve prediction precision with LSTM [13]. RNN-based models cannot consider spatial correlations in a citywide network [14].

So convolution neural network models [15] were found to focus on the space-time relationship by transforming passenger flows as images to allow the execution of convolution operations, which extracts the characteristics of the nonlinear part of the passenger flow [16]. And the residual network (ResNet) [17] has a good effect on traffic flow prediction, which is a typical framework based on the CNN layers. These models are more dependent on the design of the model structure. Hence, the prediction time increases with the complexity of the model.

Moreover, graph convolutional neural network [18], which is popular in recent years, uses graph structure data to describe the spatiotemporal relationship of traffic flow to forecast the flow [19–21]. It focuses on the relationship between macroscopic traffic flows, which is often used to forecast planning. For the trend prediction of passenger flow in the station, it needs to combine other factors (such as weather and activities) and the data of other stations to make a joint prediction to ensure the accuracy of prediction.

To improve the efficiency of the nonlinear model, the studies combine the deep-learning method with the optimization algorithm; the prediction accuracy of traffic flow is improved, such as Particle Swarm Optimization [22] and Genetic Algorithm [23]. The combination algorithm is derived, using the optimization algorithm to optimize the power value or threshold of the neural network to achieve a rapid convergence effect. At the same time, scholars hybrid some models to enhance the structural information, which will be lost during preprocessing, such as Res-LSTM [24], SVM-LSTM [25], Con-GCN [14], GA-LSTM [26], GATCN [27], PSO-LSTM [28], and TGC-LSTM [29].

However, due to the high requirement of deep learning for data, with the increase of network layers and data, it is easy to produce slow prediction speed and overfitting, which is unable to extract more accurate time-series features. A massive amount of computation cannot meet the situation of real-time and fast prediction. Generally, these deep-learning models are so complicated and they consume tremendous computing resources and training time.

The complex type of passenger flow in the comprehensive passenger terminal leads to the unobvious intention of the passenger flow line and the unstable frequent passenger flow, which is easily affected by the construction and improvement of the terminal. There are many uncertainties and nonlinear factors when predicting the passenger flow trend of integrated passenger terminals. Passenger flow data will change due to strong random factors such as weather, emergencies, and surrounding activities. Linear and nonlinear parts are intertwined, making it impossible to analyze and get more accurate prediction results.

So some scholars proposed to use the corresponding feature extraction algorithm to extract the feature of time-series data, such as EMD algorithm [27, 28], wavelet packet algorithm [30], and chaos theory [31]. Then, fusion prediction is carried out with relevant machine-learning algorithms. However, these methods need to obtain sufficient features [32]. These methods need a large amount of data support to carry out global feature analysis and extraction. Since passenger flow data is a special time-series data type, passenger flows in different periods have different features, which are not significant when global feature extraction is carried out. Hence, it is necessary to classify passenger flow patterns.

Therefore, the following problems need to be solved:

(1) How to give a fast classification of passenger flow to obtain an accurate trend of passenger flow, which can help the prediction model to learn the passenger flow trend and time-series data features.

(2) How to construct a small sample passenger flow prediction model with a simple structure, fast training speed, and good fitting effect.

As an improvement of RBF [33, 34], the GRNN neural network model [35] is similar to the RBF structure. There are more summation layers and weight connections in the hidden layer, and the output layer is removed (the least square superposition of Gaussian weight). The network generally converges to the optimal regression with a large sample concentration, which is only needed to control the SPREAD parameters for rapid prediction of a small sample size. When the sample data is small, the prediction effect of this model is better, and it can deal with unstable data. Although GRNN seems to be less accurate than radial basis, it has a great advantage in actual classification and fitting, especially when the data accuracy is relatively low.

It needs an effective method to classify and forecast the complicated passenger flow. In this paper, the passenger flow is classified by the *K*-means clustering model. Then an improved BWPs index is used to improve the clustering effect. Moreover, this GRNN neural network model is selected as the primary method of passenger flow prediction in the integrated passenger terminal, and SPREAD parameters will be adjusted by GA. The method used in our paper carried out the passenger flow prediction of different types, providing new theoretical support for the passenger flow prediction in the integrated passenger terminal.

3. An Improved *K*-Means Cluster Analysis Algorithm

As a widely used clustering algorithm, the *K*-means clustering algorithm is suitable for analyzing large sample databases. The idea of this algorithm is to calculate the physical distance between various samples and each clustering center by randomly selecting the clustering center of samples at the initial stage, then put the sample into the cluster where the clustering center is closest to it, and generate a new clustering center. Clustering calculation ends until the clustering center generated does not change. At this time, the error value reaches the minimum, and the clustering criterion function has converged.

3.1. Improved BWP_S (Between-Within Proportion-Similarity) Clustering Effect Index.

Cluster validity analysis is to evaluate the clustering effect. Generally speaking, good clustering can reflect the internal structure of the data, making the distance within the cluster as close as possible and the distance between clusters as separate as possible. The optimal cluster means that the cluster is with the smallest intracluster distance and the largest intercluster distance. Therefore, an improved clustering validity index is proposed in this section. The concept of similarity is introduced to the BWP index [36] (Between-Within Proportion) by considering the similarity of data structure, the data with high similarity within clusters can be as close as possible, and the difference of similarity between clusters can be as wide as possible.

The clustering space $b(j,i)$ represents the minimum class spacing of the i th sample of class j. The minimum class spacing means the average distance between a sample and sample groups of other categories, in which the minimum average value obtained is the minimum class spacing; $x_i^{(j)}$ represents the i th sample of class j, $x_p^{(k)}$ represents the p th sample of class k, n_k represents the number of samples of class k, and $\|\cdot\|$ represents the square Euclidean distance.

$$b(j,i) = \min_{1\le k\le c, k\ne j}\left(\frac{1}{n_k}\sum_{p=1}^{n_k}\left\|x_p^{(k)} - x_i^{(j)}\right\|^2\right). \quad (1)$$

The minimum in-class distance of the i th sample of class j is defined as $w(j,i)$, namely, the average distance between this sample and other samples of class j, and $x_q^{(j)}$ represents the q th sample of class j, $q \ne i$. n_j represents the number of samples of class j, as follows:

$$w(j,i) = \frac{1}{n_j - 1}\sum_{q=1, q\ne i}^{n_j}\left\|x_q^{(j)} - x_i^{(j)}\right\|^2. \quad (2)$$

And the $BWP(j,i)$ means the ratio of the clustering distance gap and clustering distance of the sample; that is:

$$\begin{aligned}\mathrm{BWP}(j,i) &= \frac{b(j,i)+w(j,i)}{b(j,i)-w(j,i)},\\ &= \frac{\min\limits_{1\le k\le c,k\ne j}\left(1/n_k\sum_{p=1}^{n_k}\left\|x_p^{(k)}-x_i^{(j)}\right\|^2\right)-1/n_j-1\sum_{q=1,q\ne i}^{nj}\left\|x_q^{(j)}-x_i^{(j)}\right\|^2}{\min\limits_{1\le k\le c,k\ne j}\left(1/n_k\sum_{p=1}^{n_k}\left\|x_p^{(k)}-x_i^{(j)}\right\|^2\right)+1/n_j-1\sum_{q=1,q\ne i}^{nj}\left\|x_q^{(j)}-x_i^{(j)}\right\|^2}.\end{aligned}\tag{3}$$

From the point of view of in-class compactness, the criterion to determine the validity of clustering is to make the clustering results close within the class and far away between the classes. As we can see, in terms of distance measure, the BWP index minimizes the intraclass distance and maximizes the interclass distance during clustering.

But in general, a good clustering division should reflect the internal structure of the data set as much as possible, so that the samples within the class are as similar as possible. As an index of Euclidean distance measurement, BWPs index can only get data to discuss the clustering effect from the perspective of distance, ignoring the similarity and trend among samples. Some data sets, such as gene expression data and passenger flow data, have certain sample similarities and trends, Therefore, we hope to add the concept of similarity, which can take into account the similarity measure, to maximize the intraclass similarity and minimize the interclass similarity, and then the BWPs index (Between-Within Proportion-Similarity) has been proposed. Indicators are calculated as follows.

According to Zhou's defination [37], In the clustering space, $b_s(j,i)$ is the minimum interclass dissimilarity of the i th sample of class j. The minimum dissimilarity means the average dissimilarity between a sample and sample groups of other categories, in which the minimum average value is the minimum class dissimilarity. $U(\cdot)$ represents dissimilarity.

$$b_s(j,i) = \min_{1\le k\le c,k\ne j}\left(\frac{1}{n_k}\sum_{p=1}^{n_k}U\left(x_p^{(k)},x_i^{(j)}\right)\right).\tag{4}$$

The minimum intraclass distance $ws(j,i)$ of the i th sample of class j is defined as the minimum average dissimilarity between this sample and other samples of class j, where $x_q^{(j)}$ represents the q sample of class j, and $q\ne i$, and n_j represents the number of samples in class j.

$$w_s(j,i) = \frac{1}{n_j-1}\sum_{q=1,q\ne i}^{n_j}U\left(x_q^{(j)},x_i^{(j)}\right).\tag{5}$$

Based on the similarity measure, from the perspective of in-class compactness, it is hoped that the smaller the in-class dissimilarity $w_s(j,i)$ of the sample, the better. From the perspective of the distance between classes, it is hoped that the larger the dissimilarity degree of samples from the nearest neighbor cluster, that is, the minimum dissimilarity between classes $b_s(j,i)$, the better.

Therefore, the model of the BWP index was deconstructed by referring to the BWPs index, and the similarity was discussed while considering the distance. $b_s(j,i)-w_s(j,i)$ is expressed as the cluster deviation dissimilarity to evaluate the clustering results. Obviously, the larger the $b_s(j,i)-w_s(j,i)$ is, the better the clustering effect will be.

To make the index able to analyze the validity of all samples and not be affected by dimensionality, the concept of clustering dissimilarity $b_s(j,i)+w_s(j,i)$ has been introduced. The clustering dissimilarity of a single sample is compressed through the clustering similarity of the sample so that the index becomes dimensionless, and the value of the index is the divergence dissimilarity on the clustering dissimilarity of the sample unit. Indicator values range in [−1, 1].

$\mathrm{BWP_S}$ is defined as the ratio of the clustering distance and clustering distance of the sample; that is,

$$\begin{aligned}\mathrm{BWPs}(j,i) &= \frac{b(j,i)-w(j,i)}{b(j,i)+w(j,i)} * \frac{b_s(j,i)-w_s(j,i)}{b_s(j,i)+w_s(j,i)},\\ &= \mathrm{BWP}(j,i) * \frac{\min\limits_{1\le k\le c,k\ne j}\left(1/n_k\sum_{p=1}^{n_k}U\left(x_p^{(k)},x_i^{(j)}\right)\right)-1/n_j-1\sum_{q=1,q\ne i}^{n_j}U\left(x_q^{(j)},x_i^{(j)}\right)}{\min\limits_{1\le k\le c,k\ne j}\left(1/n_k\sum_{p=1}^{n_k}U\left(x_p^{(k)},x_i^{(j)}\right)\right)+1/n_j-1\sum_{q=1,q\ne i}^{n_j}U\left(x_q^{(j)},x_i^{(j)}\right)},\end{aligned}\tag{6}$$

where $U(\cdot)$ is the dissimilarity between two samples, which can be obtained by using the Person correlation coefficient $\rho(x_i,x_k)$ of two samples as the similarity: $U(x_i,x_k)=1-\rho(x_i,x_k)$. Since the range of the correlation coefficient is [−1, 1], to ensure the nonnegativity of the function, we define the range of nonsimilarity $U(x_i,x_k)$ as [0, 2]. As shown in Figure 1, the smaller the value is, the greater the similarity is, and the larger the value is, the greater the dissimilarity is. As shown in the figure, "+" indicates a positive correlation, and "−" indicates a negative correlation. When the data exceeds 1, it indicates a negative correlation between the two data pieces.

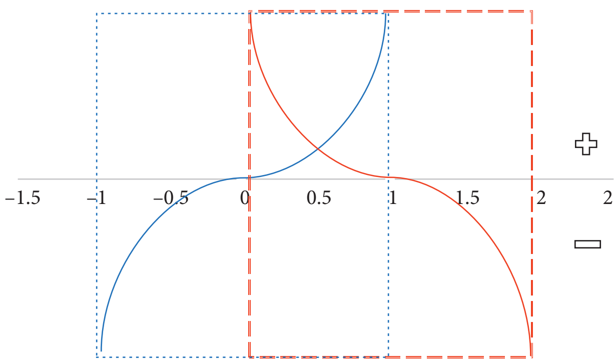


Figure 1: Diagram of dissimilarity.

$\mathrm{BWP_S}$ index can get the clustering effect of each sample in the whole dataset. As the value increases, so does the similarity of samples in the cluster. We sum the $\mathrm{BWP_S}$ values of all the samples and then average them. Therefore, the K value corresponding to the maximum value is the optimal cluster number.

$$\operatorname{avg} \mathrm{BWP}_s(k) = \frac{1}{n} \sum_{j=1}^{k} \sum_{i-1}^{n_j} \mathrm{BWP}_s(j,i), \tag{7}$$

$$k_{\mathrm{opt}} = \arg \max_{2\le k<n} \{\operatorname{avg} \mathrm{BWP}_s(k)\}. \tag{8}$$

3.2. Algorithm to Determine the Best Number of Clusters. In combination with the K-means algorithm and $\mathrm{BWP_S}$ clustering effectiveness index defined by equation (6), this section proposes an algorithm to analyze the clustering effect and determine the optimal cluster number, denoted as KMBWPS. The algorithm is summarized as follows:

(1) Select the search range of cluster number $[K_{\min}, K_{\max}]$

(2) For $K = K_{\min}$ to $K_{\max}$

①Call K-means algorithm
②Calculate the BWP index value of a single sample by using equation (6)
③Calculate the average BWP index value by using equation (7)

(3) Calculate the optimal cluster number by using equation (8)

(4) Output the best cluster number, validity index value, and clustering results

3.3. Validity Index Experiment and Analysis. To test the performance of the validity index $\mathrm{BWP_S}$ and the optimal cluster number determination algorithm, this paper carried out two groups of experiments and four data sets for testing and comparing with the common indexes: Calinski-Harabasz (CH) index [38], DaviesBouldin (DB) index [39], KrzanowskiLai (KL) index [40], Wintinter-Intra (WINT) index [41], and In-group Profit (IGP) index [42].

According to the commonly used empirical rule $K_{\max} \le \sqrt{n}$, the search range of the clustering number is $[2, K_{\max}]$. In addition, the initial clustering center is avoided to affect the clustering results of the K-means algorithm. This algorithm was run 50 times respectively to analyze the clustering effect and the generation of the optimal clustering number.

3.3.1. Experiment 1: Artificial Dataset Experiment. $\mathrm{Y_{3C}}$ data set [43] is a two-dimensional and third-class synthetic data set, which is a slightly overlapping and loose clustering structure. The optimal clustering number of the $\mathrm{Y_{3C}}$ dataset estimated by several validity indicators is shown in Table 1.

On the whole, CH index, DB index, WINT index, and $\mathrm{BWP_S}$ index all got the correct optimal clustering number, while other indexes could not get the correct optimal clustering number. From the specific situation, the CH index and $\mathrm{BWP_S}$ index have good performance, the evaluation results are stable, and the correct optimal cluster number can be obtained every time. DB index and WINT index were slightly worse, and the accuracy rates were 69% and 60%, respectively.

3.3.2. Experiment 2: UCI Real Data Set Experiment. The experiment consisted of three UCI data sets: BUPA, PIMA-Indian-Diabetes (PID), and BreastCancerWisconsin (BCW) (http://www.ics.uci.edu/mlearn/MLRepository). This $\mathrm{BWP_S}$ index was adopted to obtain the clustering evaluation results of the BUPA data set, which is based on the K-means algorithm, and the clustering evaluation results under different indexes in [36] were compared with the $\mathrm{BWP_S}$ index, as shown in Tables 2 and 3. The cluster number corresponding to the underlined index value is the best cluster number obtained from the index in this column. Since the actual class number of the BUPA data set is 2 classes, the optimal cluster number obtained is correct by CH index, DB index, IGP index, and BWPS index. In contrast, the optimal cluster number obtained is wrong by the KL index and WINT index. This indicates that the BWPS index performs well for multidimensional data.

4. GRNN Neural Network Prediction Algorithm Based on *K*-Means

The passenger flow prediction optimization algorithm is proposed, which combines the K-means and GRNN. The experiment is carried out based on the passenger flow of Chengdu East Railway comprehensive passenger terminal and compared with other prediction algorithms (GRNN, SVM, and BP).

The passenger flow data set is the daily passenger flow from January to October 2017, in which each data string is composed of the daily passenger flow with a granularity of one hour and the event label data, in the form of a vector of $[\mathrm{P}_{i1} \ldots \mathrm{P_{ij}}, \mathbf{W_i}, \mathbf{A_i}, \mathbf{M_i}]$, and $\mathrm{P_{ij}}$ represent the passenger flow at the j th hour on the i th day, and j ranges from 1 to 17. Discrete variable $\mathbf{W_i}$ represents i th day's weather which is an integer from 1 to 5 that represents the weather conditions, $\mathbf{A_i}$

Table 1: The optimal clustering number of the Y_{3C} dataset estimated by several validity indicators.

Indicators	Optimal number of clusters					Final cluster number
	2	3	4	5	Other	
CH	0	100	0	0	0	3
DB	0	69	0	0	31	3
KL	0	7	6	6	81	UNSURE
WINT	0	60	21	19	0	3
IGP	48	38	12	2	0	UNSURE
BWP_S	0	100	0	0	0	3

Table 2: Clustering validity index values of the BUPA dataset

Clustering number	CH	DB	KL	WINT	IGP	BWPs
2	**322.269**	**0.7679**	4.19	0.61	**0.985**	**0.733**
3	264.751	0.918	1.343	**0.821**	0.915	0.517
4	244.543	0.869	1.964	0.664	0.862	0.304
5	222.926	0.954	2.911	0.574	0.853	0.316
6	199.3688	0.939	0.4	0.585	0.919	0.215
7	191.999	1.016	1.971	0.546	0.845	0.207
8	181.0314	0.977	5.378	0.552	0.821	0.199
9	167.216	1.117	0.033	0.5234	0.81	0.184
10	198.317	0.933	**187.678**	0.537	0.851	0.2178
11	184.512	0.979	0.047	0.521	0.864	0.185
12	178.436	0.963	2.081	0.537	0.843	0.192

Table 3: The best clustering number of the UCI data set estimated by several validity evaluation indexes [36].

Date set	Sample dimension	BWP	Real value	CH	DB	KL	WINT	IGP	BWPs
BUPA	6	2	2	2	2	10	3	2	2
PID	8	2	2	3	4	7	3	2	2
BCW	9	2	2	2	2	3	3	2	2

is a 0–1 variable which means whether there is activity around the i th day, and $\mathbf{M_i}$ is also a 0–1 variable which indicates whether there is an emergency on the i th day.

The main ideas of the algorithm are as follows: firstly, the obtained passenger flow data are processed, and abnormal data are screened and eliminated. Second, K-means clustering algorithm is used to group the processed data, and BWPs index is used for optimal selection. Then, selecting K value at the maximum value of BWPs index and obtaining the best effect of clustering, to get the best clustering groups, the GRNN neural network is used to train the clusters, and the corresponding neural network model is obtained. Finally, comparing the similarity between the predicted data and K clustering centers, the neural network model with a high matching degree was found for prediction. This algorithm steps are as follows.

Step1. Data analysis and data preprocessing

4.1. Data Analysis. According to the passenger flow data of Chengdu East integrated passenger hub metro system from January to October 2017, this paper takes passenger flow of 17 hours a day as a set of data. These time series can be processed into several different categories because of nonlinear factors (e.g., weather changes, emergencies).

4.2. Data Preprocessing. These samples can be taken at intervals of one hour from 6:00 a.m. to 11:00 p.m. and 17 sampling points can be obtained per day. After the abnormal data and data fault days caused by emergencies (no data was generated due to the shutdown of the gate during the period of large passenger flow due to emergencies) were eliminated, finally, 298 days of valid data were obtained. Among them, 80% of passenger flow data is taken as model test data and 20% as validation data. At the same time, cross-validation is carried out to ensure the validity of the model.

Step2. K-means clustering

The selected 80% passenger flow data was performed using K-means clustering, and the selection range of K value is 2 to 18, and it is checked ten times for each K value. Then, the clustering situation with the maximum BWPs index was taken as the optimal clustering effect of the current K value. Finally, the K value which gets the highest BWPs index was selected as the optimal K value, and K-means clustering and grouping are performed.

Step3. Prediction based on similarity analysis

According to the above calculation, the K clustering centers can be obtained with K passenger flow patterns. At the same time, comparing the similarity between the initial sampling time data of K clustering centers and the initial sampling time data of the day to be predicted, the nearest group of data is taken as the GRNN neural network test sample, and the GRNN neural network model of this group of data is established. Then, the propagation velocity of the radial basis is optimized by a genetic algorithm. Finally, the prediction is made according to the forecast data.

5. The Example Analysis

5.1. K-Means Clustering under BWPs Index. *K*-means clustering, grouping, and BWPs index calculation were carried out for passenger flow data to obtain BWPs index under different *K* values, as shown in Table 4 below. It can be seen from the table that when the *K* value is 16, the BWPs index is the maximum. Hence, when the *K* value is 16, the best clustering effect can be obtained. Sixteen passenger flow patterns were finally obtained, as shown in Figure 2.

To analyze these 16 passenger flow modes, they were converted into passenger flow heat maps. The horizontal axis was the time axis, and the vertical axis was different passenger flow modes. The color change represented the change of passenger flow, and the darker the color was, the greater the passenger flow was, which has taken 10000 as the limit passenger flow. As shown in Figure 3, type2 and type8 have the most obvious fluctuation of passenger flow; type1's passenger flow is relatively stable and has no large fluctuation change. For all passenger flow modes, there is a significant drop in passenger flow and it formed a trough at sampling point 7 (12:00 p.m. to 1:00 p.m.), and it is easy to generate wave peaks at sampling points 6, 8, 11, and 15.

The subway passenger flow of Chengdu East Railway Station is composed of railway passenger flow and commuter passenger flow of surrounding business districts, which means the trend of the passenger flow easily receives the influence of railway trains arrangement (for station passenger flow, passengers arrival time is relatively fixed). At the same time, the generation of wave crest is related to the number of vehicles arriving at the sampling point, and the number of vehicles in this period increases compared with other sampling points.

At the same time, the generation of wave crest is related to the number of arriving trains at this sampling point. The number of arriving vehicles at this sampling point increases compared with other sampling points, and the passenger flow will also increase. The valley value at sampling point 7 is because the number of arriving trains decreases at this sampling point and there is less commuter traffic during this period. In order to better analyze the trend of passenger flow, sixteen passenger flow patterns are divided into five categories by similarity measurement, as shown in Figure 4.

As shown in Figure 4(a), the passenger flow trend of type1 is very distinct from that of type2 and type8. Type1 belongs to the peak passenger flow, passenger flow fluctuation is not obvious throughout the day, according to the label. This type of passenger flow is mainly concentrated in the middle of the Spring Festival when social activities are reduced. It can be seen that most of the passenger flow is commuter flow, and the railway passenger flow accounts for a small proportion by querying the passenger flow through security check on that day (railway passenger flow does not need security check to enter the subway, so the passenger flow proportion can be obtained).

Table 4: BWP$_S$ indicators.

K Value	2	3	4	5	6	7	8	9	10
BWPs	0.41	0.47	0.63	0.75	0.63	0.77	0.82	0.80	0.84
K Value	11	12	13	14	15	16	17	18	
BWPs	0.67	0.8	0.83	0.73	0.82	0.91	0.84	0.86	

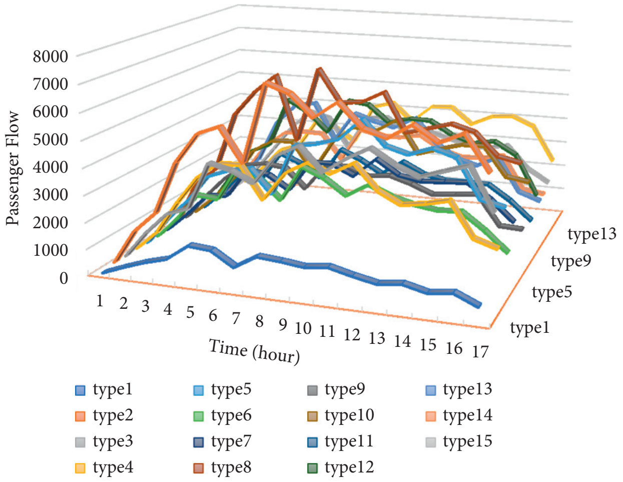


Figure 2: Passenger flow pattern.

Type2 and type8 are typical bimodal passenger flows, and their peaks are generated at sampling points no. 6 and no. 8, respectively. The passenger flow has always been at a high level, which is easy to cause the accumulation of passenger flow. According to the labels, it can be found that this kind of passenger flow mainly comes from the Spring Festival passenger flow, which occurs in the peak period of returning to hometown and returning to the city. The purpose of passenger flow is relatively clear and not easy to be affected by the weather and surrounding activities, mainly for returning home for holidays and leaving home for work.

As shown in Figure 4(b), the passenger flow pattern in the graph has a multipeak feature, and the crest stays longer and falls slower. According to the security check information on that day, 80% of the passenger flow comes from railway passenger flow, and most of them are medium-long distance passenger flow, which is greatly affected by railway passenger flow. The passenger flow label shows that there will be local influential activities in Chengdu around that day, such as expos, sugar, and wine fairs. The passenger flow shows a certain trend with the activity size and type5 has extreme weather in the afternoon, resulting in a surge of passenger flow.

As shown in Figure 4(c), it is a typical passenger flow of 4 peaks, and the peak is mainly concentrated at the commuting time points. The passenger flow reaches the peak at 9 o'clock, 11 o'clock, 2 o'clock, and 5 o' clock. The passenger

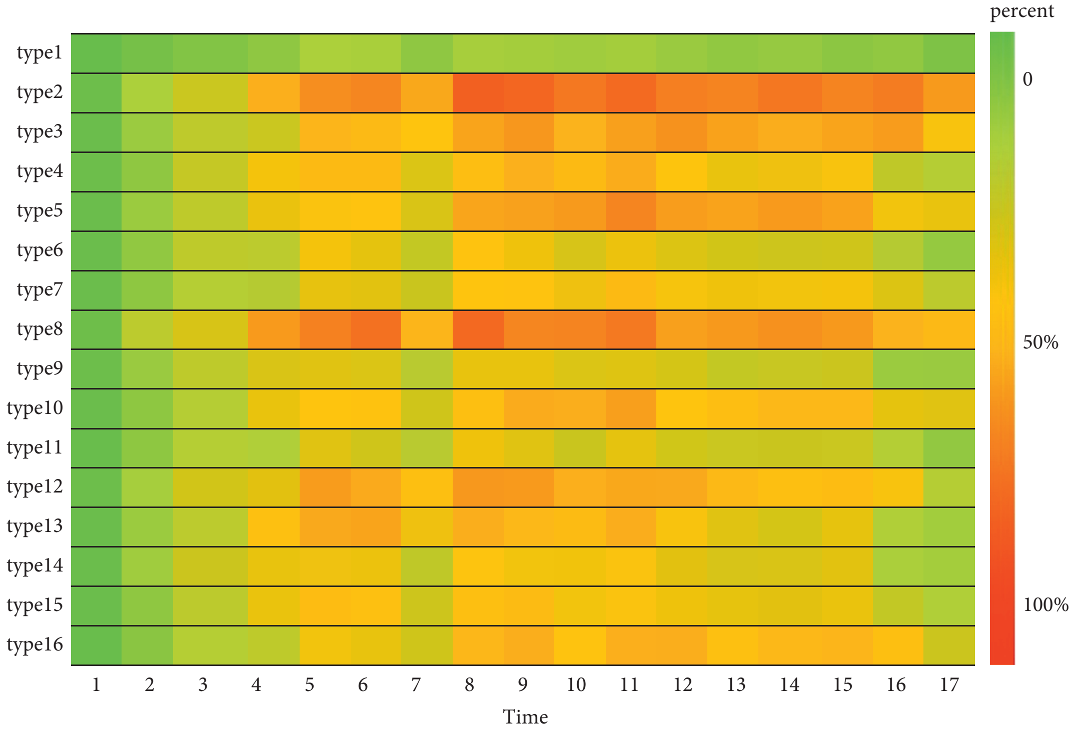
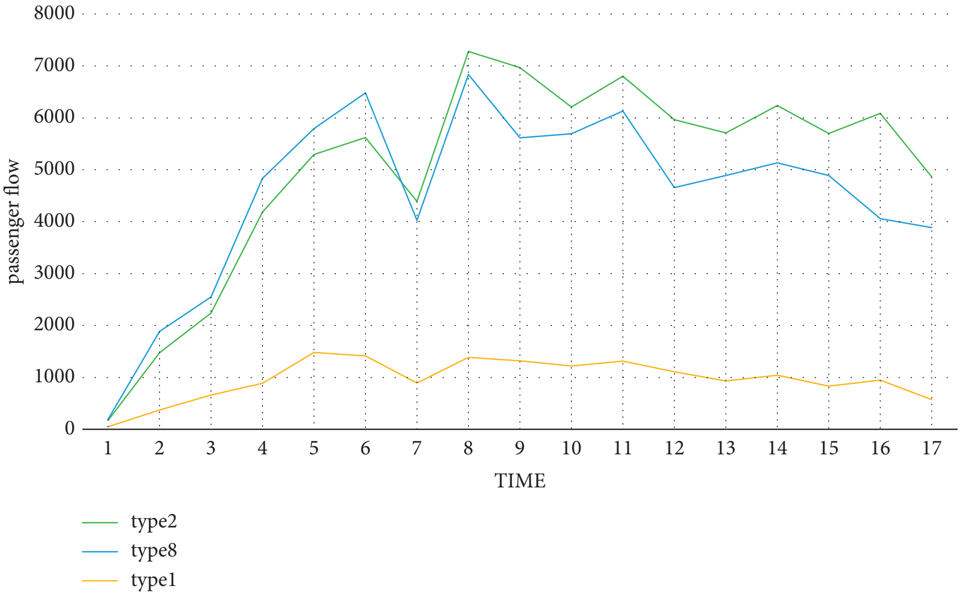
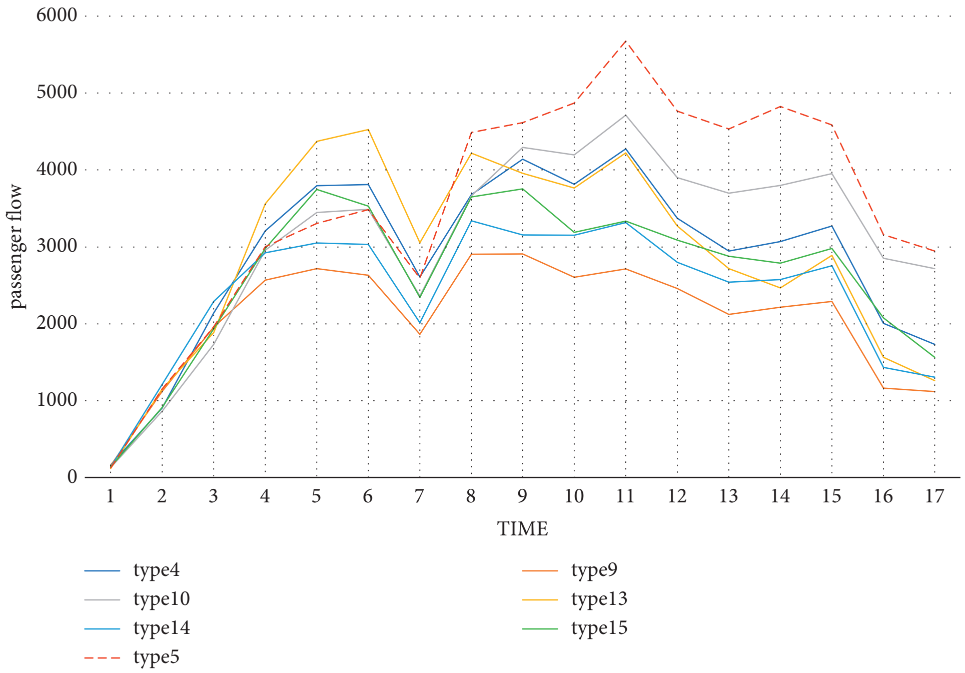


Figure 3: Passenger flow pattern thermodynamic diagram.

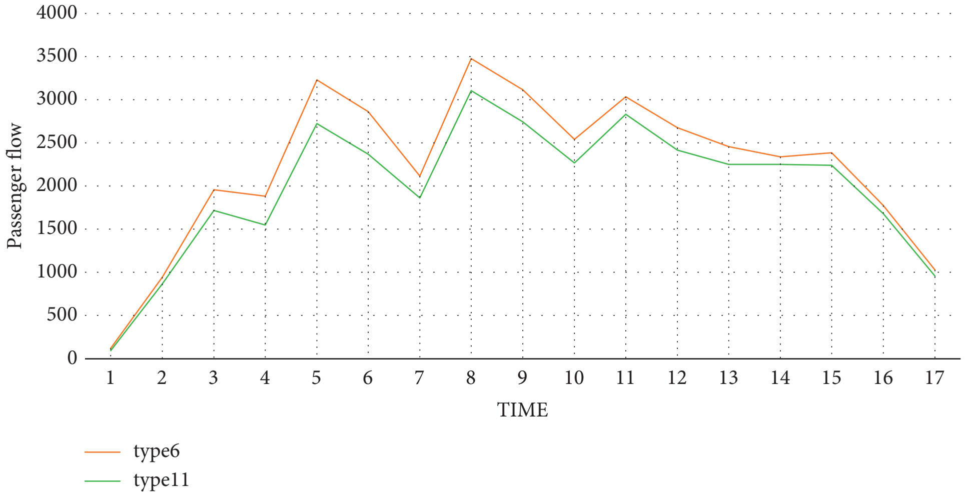


(a)

Figure 4: Continued.

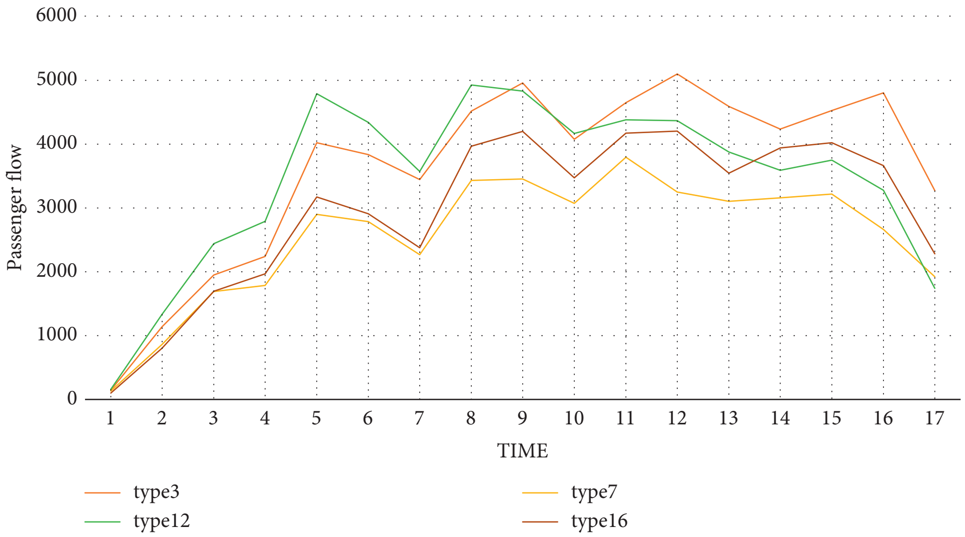


(b)



(c)

Figure 4: Continued.



(d)

Figure 4: Trend of passenger flow type.

flow is mainly from Monday to Friday, without activities or holidays in labels, called commuter flow patterns.

As shown in Figure 4(d), the wave peak lasts for a certain period, and the decline is slow. The situation is similar to the passenger flow of Spring Festival, but the volume is relatively small, 80% of which is commuter passenger flow. This type of passenger flow mainly occurs around the regular holidays, called the holiday passenger flow pattern.

5.2. GRNN Neural Network Prediction Model Test. According to the Granger causality, it is found that there is a Grange causality between the passenger flow in the first 5 hours and the passenger flow in the sixth hour. Therefore, when establishing the prediction model, the first five passenger flows of the historical data are taken as input and the sixth passenger flow is taken as output, and data training is conducted respectively in the form of a sliding window, which will train 12 GRNN neural networks. As shown in Figure 5, the input is $[x_i, x_{i+1}, x_{i+2}, x_{i+3}, x_{i+4}]$ which consists of Y_i and X_{i+4}, the output is x_{i+5}, Y_i is a vector form $[x_i, x_{i+1}, x_{i+2}, x_{i+3}]$, and i is range from 1 to 12.

Therefore, we compare the similarity between the first 5 sampling points of the day to be predicted and the first 5 sampling points of the 16 passenger flow patterns and select the group with the highest similarity for neural network training.

At the same time, the first 5 sampling points in a day are used to predict the passenger flow data of the following 11 sampling points by using the way of sliding window (as shown in Figure 6), which means that the input layer has 5 neurons (the passenger flow at the first 5 sampling moments) and the output layer contains 1 neuron. SVM and BP reference this way for prediction comparison.

GRNN has four layers of network structure, including input layer, pattern layer, summation layer, and output layer. The error of output data and training samples of the GRNN neural network is mainly determined by the smoothing factor. Therefore, GRNN neural network has a very simple performance control mode, which can obtain better performance only by adjusting the smoothing factor [35].

In the passenger flow prediction model, each GRNN is treated as a block. In the $\mathrm{GRNN_i}$ model, Y_i and X_{i+4} have been taken as inputs and X_{i+5} as outputs. At the same time, the output value and $Yi+1$ are used as the input of the next GRNN$i+1$ model, and so on. A recurrent neural network structure is formed, which can be segmented to extract features of different periods.

The expansion speed of the radial basis can be obtained through the genetic algorithm, and its accuracy can be verified. In Figure 7, the X-axis represents 17 sampling points on July 1st, and the sampling points data belongs to the test data. The Y-axis represents the ratio of the hourly passenger flow to the maximum passenger flow on that day. As seen in Figure 7, in general, the prediction error of the GRNN neural network model based on K-means clustering is small. This effect of the base test is better than that of BP and SVM. And the network output is in better agreement with the expected response. The training time of GRNN is 27 percent and 43 percent less than that of SVM and BP, respectively.

5.3. GRNN Neural Network Prediction Model Validation. Therefore, a similarity comparison was conducted between the first 5 sampling points of passenger flow data on November 1, 2017, which belongs to validation data, the day is selected randomly, and 16 passenger flow models are in the model to select passenger flow models with high similarity. Neural network training and prediction were carried out following Step 2, and cross-validation was carried out. The results are shown in Table 5 and Figure 8.

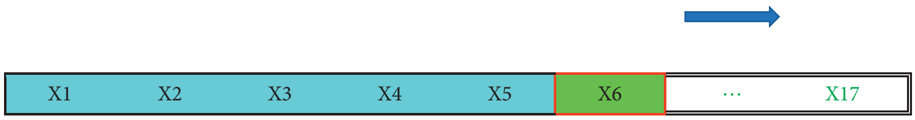


Figure 5: Passenger flow forecasting data input and output.

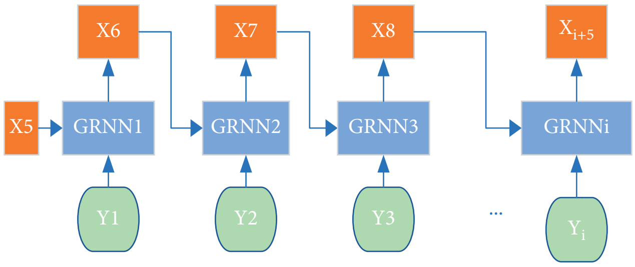


Figure 6: Passenger flow prediction model.

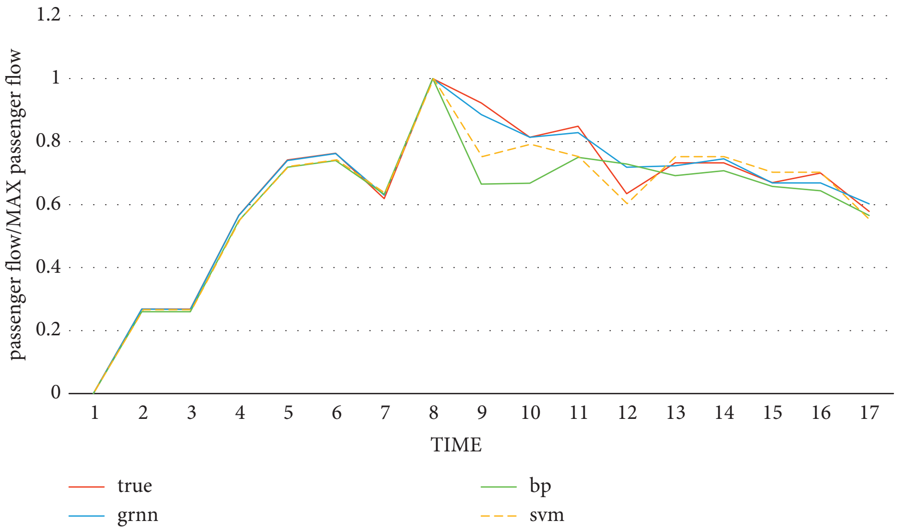


Figure 7: Test comparison of different prediction methods (July 1).

Table 5: Validation prediction and evaluation indexes.

The evaluation index	MSE	RMSE	MAE	MAPE	R2
K-means-GRNN	0.0083	0.0912	0.101	0.1356	0.799
K-means-BP	0.0336	0.1833	0.211	0.2774	0.191
K-means-SVM	0.0177	0.133	0.241	0.283	0.691

According to the test results in Table 5 and Figure 8, the prediction error of the validation in the GRNN neural network model after clustering is less than that of the BP neural network, the predicted value of the SVM and BP model is generally higher than the true value, and the predicted value under the GRNN model better reflects the trend of the true value. Based on the cross-validation of test data and inspection data, it is proved that the expected response of the GRNN neural network model is better than that of the SVM and the BP neural network.

Finally, we can draw some conclusions as follows. The passenger flow pattern of the station can be divided into 16 types. Compared with other prediction models, the neural network prediction based on *K*-means is closer to the actual value.

Moreover, in most cases, the prediction model of GRNN based on *K*-means clustering has advantages in both speed and accuracy. Therefore, the validity of GRNN based on *K*-means clustering can be verified by the above experiments. This algorithm can be applied to predict the trend of large

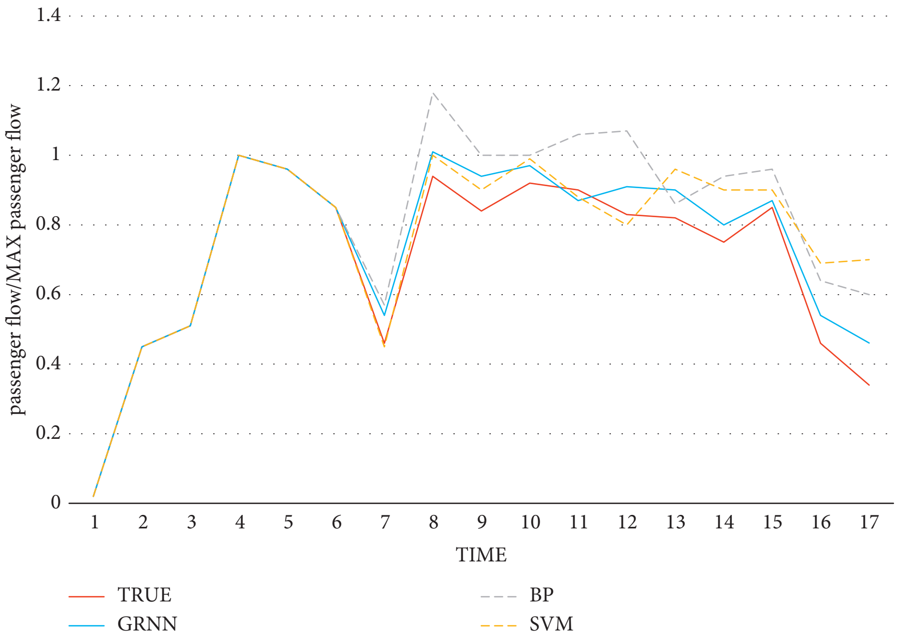


FIGURE 8: Validation comparison of different prediction methods.

passenger flow, which is beneficial to the early warning of the comprehensive passenger terminal.

6. Conclusions

K-means-GRNN neural network model is proposed in this paper, which improves the correlation of data through the BWPs index and selects the data group with high matching to predict the passenger flow trend through GRNN. The BWPs index can take into account the similarity measure, to maximize the intraclass similarity and minimize the interclass similarity, which can be used to classify passenger flows with the same trend characteristics to facilitate the extraction of time-series features later.

In the prediction model, a recurrent neural network structure is constructed by 12 blocks (GRNN), and features of corresponding periods are extracted by each block. This model reduces the problem of network complexity and a large amount of computation. The main findings of the study are summarized as follows:

(1) It is found that this model can better improve the clustering effect of *K*-means, comparing the improved $\mathrm{BWP_S}$ index with several other classical indexes.

(2) The passenger flow of Chengdu East Railway Station has been divided into 16 models, which have been classified as 5 trends such as activity passenger flow, regular commuter passenger flow, and holiday passenger flow.

(3) Experimental results show that the prediction accuracy of the GRNN neural network based on *K*-means clustering is higher than that of the BP network model and SVM model, and the error is small.

However, in the passenger flow prediction, the randomness of *K*-means clustering based on indicators is very strong, and the appropriate clustering combination cannot be obtained in a short time. Therefore, future studies should consider improving the clustering efficiency to obtain more accurate prediction efficiency.

Data Availability

The passenger flow data were provided by Chengdu Metro Group and Chengdu Municipal Government. The data provided are for use only in this paper because it involves commercial and government requirements for the use of data.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

This work was supported by Chengdu Science and Technology Bureau Science and Technology Benefit Project (2015-HM01-00043-SF); National Key Research and Development Program of China (2017YFB1200703); and National Natural Science Foundation of China (51278429).

References

[1] Z. Yao, T. Xu, Y. Jiang, and R. Hu, "Linear stability analysis of heterogeneous traffic flow considering degradations of connected automated vehicles and reaction time," *Physica A: Statistical Mechanics and Its Applications*, vol. 561, Article ID 125218, 2021.

[2] H. Ren, Y. Song, and S. Li, "Two-stage optimization of Urban rail transit formation and real-time station control at comprehensive transportation hub," 2020, https://arxiv.org/abs/2008.12207.

[3] C. Yu, H. Li, X. Xu, and J. Liu, "Data-driven approach for solving the route choice problem with traveling backward behavior in congested metro systems," *Transportation Research Part E: Logistics and Transportation Review*, vol. 142, Article ID 102037, 2020.

[4] X. Zhang, X. Guo, D. U. Xiaochuan, and R. Deng, "The planning points and guidelines of the comprehensive passenger transport hub," *Modern Urban Research*, no. 10, pp. 115–120, 2013.

[5] C. M. Hurvich and C. Tsai, *Regression and Time Series Model Selection*, World Scientific Publishing Company, Singapore, Singapore, 1998.

[6] B. L. Smith and M. J. Demetsky, "Traffic flow forecasting: comparison of modeling approaches," *Journal of Transportation Engineering*, vol. 123, no. 4, 1997.

[7] S. Shekhar and B. Williams, "Adaptive seasonal time series models for forecasting short-term traffic flow," *Transportation Research Record Journal of the Transportation Research Board*, vol. 2024, no. 2024, pp. 116–125, 2007.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via theEMAlgorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[9] G. Van, J. A. K. Suykens, D.-E. Baestaens et al., "Financial time series prediction using least squares support vector machines within the evidence framework," *IEEE Transactions on Neural Networks*, vol. 12, no. 4, pp. 809–821, 2001.

[10] K. Kumar, M. Parida, and V. K. Katiyar, "Short term traffic flow prediction for a non urban highway using artificial neural network," *Procedia—Social and Behavioral Sciences*, vol. 104, pp. 755–764, 2013.

[11] X. Ma, Z. Tao, Y. Wang, H. Yu, and Y. Wang, "Long short-term memory neural network for traffic speed prediction using remote microwave sensor data," *Transportation Research Part C: Emerging Technologies*, vol. 54, pp. 187–197, 2015.

[12] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," *Neural Computation*, vol. 1, no. 2, pp. 270–280, 1998.

[13] D. Yang, K. Chen, M. Yang, and X. Zhao, "Urban rail transit passenger flow forecast based on LSTM with enhanced long-term features," *IET Intelligent Transport Systems*, vol. 13, no. 10, pp. 1475–1482, 2019.

[14] J. Zhang, F. Chen, Y. Guo, and X. Li, "Multi-graph convolutional network for short-term passenger flow forecasting in urban rail transit," *IET Intelligent Transport Systems*, vol. 14, no. 10, pp. 1210–1217, 2020.

[15] S. Guo, Y. Lin, S. Li, Z. Chen, and H. Wan, "Deep spatial-temporal 3D convolutional neural networks for traffic data forecasting," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3913–3926, 2019.

[16] X. Ma, Z. Dai, Z. He, J. Ma, Y. Wang, and Y. Wang, "Learning traffic as images: a deep convolutional neural network for large-scale transportation network speed prediction," *Sensors*, vol. 17, no. 4, p. 818, 2017.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, June 2016.

[18] T. N. Kip F and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, https://arxiv.org/abs/1609.02907.

[19] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Graph convolutional recurrent neural network: data-driven traffic forecasting," 2017, https://arxiv.org/abs/1707.01926.

[20] W. Jiang and J. Luo, "Graph neural network for traffic forecasting: a survey," 2021, https://arxiv.org/abs/2101.11174.

[21] L. I. Zhishuai, L. Yisheng, and G. Xiong, "Short-term traffic flow prediction based on graph convolutional neural network and attention mechanism," *Journal of Transportation Engineering*, vol. 19, no. 4, pp. 15–19+28, 2019.

[22] J. L. F. Martínez, "A brief historical review of particle swarm optimization (PSO)," *Journal of Bioinformatics and Intelligent Control*, vol. 1, no. 1, pp. 3–16, 2012.

[23] D. Whitely, "Genetic algorithm and neural networks: optimizing connexions and connectivity," *Parallel Computing*, vol. 14, 1990.

[24] J. Zhang, F. Chen, Y. Zhu, and Y. Guo: Deep-learning architecture for short-term passenger flow forecasting in urban rail transit', 2019, http://www.researchgate.net/publication/338292639_Deep-learning_Architecture_for_Short-term_Passenger_Flow_Forecasting_in_Urban_Rail_Transit.

[25] J. Guo, Z. Xie, Y. Qin, L. Jia, and Y. Wang, "Short-term abnormal passenger flow prediction based on the fusion of SVR and LSTM," *IEEE Access*, vol. 7, pp. 42946–42955, 2019.

[26] H. Wen, D. Zhang, and L. U. Siyuan, "Application of GA-LSTM model in highway traffic flow prediction," *Journal of Harbin Institute of Technology*, vol. 51, no. 9, pp. 81–87+95, 2019.

[27] G. Guo and W. Yuan, "Short-term traffic speed forecasting based on graph attention temporal convolutional networks," *Neurocomputing*, vol. 410, pp. 387–393, 2020.

[28] Y.-Y. Qiu, Q. Zhang, and M. Lei, "Forecasting the railway freight volume in China based on combined PSO-LSTM model," *Journal of Physics: Conference Series*, vol. 1651, no. 1, Article ID 012029, 2020.

[29] Z. Cui, K. Henrickson, R. Ke, and Y. Wang, "Traffic graph convolutional recurrent neural network: a deep learning framework for network-scale traffic learning and forecasting," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 11, pp. 4883–4894, 2020.

[30] H. Huang, J. Chen, X. Huo, Y. Qiao, and L. Ma, "Effect of multi-scale decomposition on performance of neural networks in short-term traffic flow prediction," *IEEE Access*, vol. 99, pp. 50994–51004, 2021.

[31] A. Cheng, X. Jiang, Y. Li, C. Zhang, and H. Zhu, "Multiple sources and multiple measures based traffic flow prediction using the chaos theory and support vector regression method," *Physica A: Statistical Mechanics and Its Applications*, vol. 466, pp. 422–434, 2017.

[32] A. Miglani and N. Kumar, "Deep learning models for traffic flow prediction in autonomous vehicles: a review, solutions, and challenges," *Vehicular Communications*, vol. 20, pp. 100184.1–100184.36, 2019.

[33] P. I. Jun, M. A. Sheng, Q. Zhang, L. Wang, and D. Cui, "Aero-engine exhaust gas temperature prediction model based on IFOA-GRNN," *Journal of Aerospace Power*, vol. 34, no. 1, pp. 8–17, 2019.

[34] A. Hl, A. Yw, A. Xx, B. Lq, and A. Hz, "Short-term passenger flow prediction under passenger flow control using a dynamic radial basis function network," *Applied Soft Computing*, vol. 83, Article ID 105620, 2019.

[35] D. F. Specht, "A general regression neural network," *IEEE Transactions on Neural Networks*, vol. 2, no. 6, pp. 568–576, 1991.

[36] S. Zhou, Z. Xu, and X. Tang, “Method for determining optimal number of clusters in K-means clustering algorithm,” *Journal of Computer Applications*, vol. 30, no. 8, pp. 1995–1998, 2010.

[37] S. Zhou, *Research and application on determining optimal number of clusters in cluster analysis*, Ph.D. thesis, Jiangnan University, Wuxi, China, 2012.

[38] T. Calinski and J. Harabasz, “A dendrite method for cluster analysis,” *Communications in Statistics - Simulation and Computation*, vol. 3, no. 1, pp. 1–27, 1974.

[39] D. L. Davies and D. W. Bouldin, “A cluster separation measure,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979.

[40] S. Dudoit and J. Fridlyand, “A prediction-based resampling method for estimating the number of clusters in a dataset,” *Genome Biology*, vol. 3, no. 7, 2002.

[41] E. Dimitriadou, S. Dolničar, and A. Weingessel, “An examination of indexes for determining the number of clusters in binary data sets,” *Psychometrika*, vol. 67, no. 1, pp. 137–159, 2002.

[42] A. V. Kapp and R. Tibshirani, “Are clusters found in one dataset present in another dataset?” *Biostatistics*, vol. 8, no. 1, pp. 9–31, 2007.

[43] D. Dembele and P. Kastner, “Fuzzy C-means method for clustering microarray data,” *Bioinformatics*, vol. 19, no. 8, pp. 973–980, 2003.