Research on Sample Selection of Urban Rail Transit Passenger Flow Forecasting Based on SCBP Algorithm

WENBO LU, CHAOQUN MA, AND PEIKUN LI
College of Transportation Engineering, Chang’an University, Xi’an 710064, China
Corresponding author: Chaoqun Ma (machaoqun@chd.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 71871027, and in part by the Ministry of Housing and Urban-Rural Development under Grant 2016-K2-033.

ABSTRACT Due to the wide applications of deep learning in the field of urban rail transit passenger flow forecasting, the selection problem of training samples has become increasingly more worthy of researchers’ attention, as it is closely related to urban rail transit passenger flow time series. Therefore, it is necessary to study the distribution characteristics of the contribution degree of the training sample to guide sample selection in the deep learning training process. In this study, based on the prediction accuracy and the sample contribution degree, the optimal sample contribution combination algorithm (GWO-SCBP) was ultimately generated by the grey wolf optimizer (GWO) and error back propagation (EBP) algorithms. The contribution of training samples for each station of the Xi’an metro network was calculated and analyzed. The results show that the sample contribution is not only related to the distance between the sample and predicted value, but is also closely related to the station flow characteristics. By classifying the network stations and fitting the contribution degree of the central station of each type of station, linear equations of sample contribution degree were obtained, and the $R^2$ values attained at least 0.65, indicating a good fitting effect.

INDEX TERMS Rail transit, passenger flow forecast, sample selection, grey wolf optimizer algorithm, artificial neural network.

I. INTRODUCTION The demand for public transportation has become increasingly greater with the expansion of cities, and a unique method of urban rail transit operation is required to adapt to city development. The passenger flow of urban rail transit determines the scale of the station, the time of train operation, and the train formation. A reasonable urban rail transit operation organization can not only increase the safety and punctuality of travelers but also reduce the travel time. Therefore, it is of great significance for urban rail transit to determine the scale of passenger flow in the future. Urban rail transit passenger flow forecasting can be mainly categorized into near, early-stage, and long-term forecasting. Long-term and early-stage forecasting primarily rely on the typical four-stage method for prediction. In contrast, near passenger flow forecasting includes time series forecasting, time feature analysis, and regression analysis. Urban rail transit passenger flow is characterized by a long-term linear growth trend, cyclical seasonal variation, and random fluctuation and thereby exhibits complex nonlinear characteristics [1]. Many methods have been developed to improve passenger flow prediction accuracy and these methods can be primarily divided into two categories: parametric methods and nonparametric methods. The Autoregressive Integrated Moving Average model (ARIMA) [2] is widely regarded as the classical parametric method. Milenkovic et al. [3] added seasonal changes in passenger flow to the ARIMA model to overcome the impact of passenger flow fluctuations with regular cycles. Parametric methods can achieve good performance and capture regular changes well, but cannot cope with the irregular changes of urban rail transit passenger flow. Thus, nonparametric methods, such as support vector machines (SVMs), long short-term memory neural network (LSTM), and so on, have been utilized to overcome this problem, and have achieved great success. Tang et al. [7] successfully added spatial variables to traditional time series prediction...
and applied it to LSTM networks. Qin et al. [8] used the Echo State Network (ESN) to predict the passenger flow of each part of seasonal decomposition. As well as, many studies about the rail transit passenger flow forecasting of events have got great progress [9], [10].

However, these methods do not take into account sample selection when making predictions, although existing research has proven that sample selection has a great impact on trained network models [11]–[14]. In addition, many studies have confirmed that appropriate sample selection can increase model accuracy [15], [16].

Urban rail transit passenger flow forecasting has strong requirements for sample integrity, such as the consideration of special periods, like holidays. To investigate the problem of the uneven distribution of training samples, the concept of sample contribution degree is proposed in this paper, the error back propagation (EBP) algorithm is deduced, and the EBP algorithm based on sample contribution degree (SCBP) is ultimately obtained. Based on the optimization strategy of the grey wolf optimizer (GWO) algorithm, the distribution of the sample contribution of each station is found on the premise of predicting error. The sample contribution of each station is analyzed, and the influencing factors of the sample contribution degree are obtained, which provides a reference for the sample selection of the deep learning prediction method in the field of urban rail transit passenger flow forecasting.

The remainder of this paper is organized as follows. A review of related work on data selection and sample selection is provided in Section 2. Section 3 introduces the concept of sample contribution, the principle of the EBP algorithm, the EBP algorithm combined with sample contribution, and the framework for determining sample contribution. In Section 4, taking the Xi’an rail transit out-flow forecast as an example, a feasible sample contribution optimization framework for this example is designed, and the complete process is demonstrated. In Section 5, the optimization result is verified and factors that affect the sample contribution in aspect of the training sample features and the prediction sample features are analyzed. In Section 6, stations are classified by the influencing factors of sample contribution, various types of central stations are analyzed, and specific fitting is conducted. Finally, the conclusions are drawn in Section 7.

II. RELATED WORK
The uncertainty of research object selection directly affects the accuracy of passenger flow prediction. Before training model, the data need to be organized, such as time granularity, time interval, etc. In the model training process, we also need to pay attention to the training sample selection. Therefore, considerable research efforts have been made to investigate the influence of prediction accuracy due to the data selection and sample selection in recent decades. According to its steps, we divided it into the data selection and sample selection.

A. THE DATA SELECTION WORK
Many scholars have done a lot of research in terms of data selection. The lack of selection criteria for passenger flow prediction data will cause large fluctuations in prediction accuracy. Zhang et al. [17] proposed to generate multi-day activity-travel data through sampling from readily available single-day household travel survey data. Li et al. [18] indicated that the time granularity of the prior passenger flow data and the difference in site classification will also lead to differences in prediction accuracy. Ma et al. [19] explored the distribution of passenger spatiotemporal characteristics based on big data of traffic smart cards, laying a foundation for the selection of passenger flow prediction objects. Chiu and Verma [20] used the method of data size to divide the data into different groups, and conducting training and testing of neural network ensembles.

B. THE SAMPLE SELECTION WORK
Existing research has proven that sample selection has a great impact on trained network models. The traditional focus has been on active learning and information compression.

Active learning is the use of the internal correlation of the sample to learn new samples autonomously, and some representative examples from the unmarked collection are determined, thus reducing the cost of marking. It generally contains three types of queries: uncertainty-based queries [21]–[24], version space-based queries [25], [26], and expected error-based queries [27], [28]. Tong et al. [29] accessed many useless examples and autonomously found the needed examples, thereby reducing the need for labeled examples. Melville and Mooney [30] used committees produced by Decorate to select examples for labeling. Nguyen and Smeulders [31] proposed a formal model for the incorporation of clustering into active learning.

Information compression is designed to reduce training time by filtering noise and redundant data to compress data sets. The representative algorithms of information compression are the condensed nearest neighbor (CNN) rule [32], [33], and principal component analysis (PCA) [34], [35], among others [36]. Nikolaidis et al. [37] introduced a multi-stage method for pruning the training set, which can much improve storage reduction and competitive execution speeds. Yang et al. [38] presented a novel editing algorithm called adaptive Edited Natural Neighbor algorithm (EaN), which can improve the performance of condensation methods in terms of both accuracy and reduction rate greatly. Tsai et al. [39] presented a novel undersampling approach that combines clustering analysis and instance selection. The clustering analysis component groups similar data samples of the majority class dataset into 'subclasses', while the instance selection component filters out unrepresentative data samples from each of the 'subclasses'.

This paper aims to sample selection. Compared with active learning and information compression, we have the following differences: 1. The number of training samples was not reduced, which can improve the utilization rate of the sample. 2. We proposed the concept of sample contribution and applied it to the field of deep learning. The training sample contribution is given by the optimization algorithm in the
training process. 3. Our proposed SCBP algorithm can be applied to all artificial neural network models based on the EBP algorithm.

The innovations of this paper are as follows: (i) the concept of the sample contribution is put forward and integrated into the process of adjusting weights by the classical EBP algorithm; (ii) a feasible optimization framework based on the intelligent optimization algorithm is proposed; (iii) taking the passenger flow forecasting of urban rail transit as an example, characteristics of passenger flow are analyzed, and a hierarchical method for optimizing sample contribution is proposed; (iv) the proposed model is highly extensible in other networks based on the EBP algorithm.

III. METHODS
Sample selection is a core issue in the process of network training. Depending on the size and attributes of the samples, the effects of neural network training can vary greatly. Existing research has indicated that the impact of training samples on training outcomes is different. Assuming that the contribution of each training sample to the training effect can be calculated, then during the process of training, a better neural network model can be obtained by adjusting the weight of the network according to the contribution degree. In this section, the determination of how to find the contribution of each training sample from the existing training sample and forecasted sample is presented in detail.

First, the details of the concept of sample contribution are discussed. Then, by analyzing the principle of the classical error back propagation algorithm, the contribution of each training sample is added, and the error adjustment strategy based on the sample contribution degree is obtained. Next, the more commonly used GWO intelligent optimization algorithm is introduced. Finally, after integrating the sample contribution into the EBP algorithm, the GWO algorithm is used to optimize each training sample contribution, which can be applied to future training to obtain a better neural network prediction model.

A. SAMPLE CONTRIBUTION
Contribution is an indicator of economic efficiency, and refers to the ratio of the number of effective or useful results to the consumption and occupancy of resources, i.e., the ratio of output to input, or the ratio of the amount obtained to the total amount. Contribution is also used to analyze the extent of the effects of various factors in economic growth. The 80/20 rule (Pareto principle) [40] was put forward by the Italian economist Pareto in the late 19th and early 20th centuries. He believed that in any group of things, the most important part is only a small portion of about 20%, and the remaining 80%, although the majority, is secondary. The 80/20 rule states that there is an unexplained imbalance between cause and effect. In general, inputs and efforts can be categorized into two different types: a majority, which only has a small impact, or a few, which has major and significant impacts.

By applying this rule to sample training, it can be concluded that in the sample training process, 80% of attention is placed on 20% of the sample, and 20% of attention is placed on 80% of the sample. A better prediction model may thus be obtained. Therefore, the sample contribution degree can be defined as the guiding effect of the training process on the adjustment error. When the contribution degree of the sample is larger, the direction of the adjustment error is closer to this sample. Assuming that the number of training samples is n, the contribution of each sample is $c_i$, and $\sum_{i=1}^{n} c_i = 1$.

B. EBP ALGORITHM BASED ON SAMPLE CONTRIBUTION
In this section, the EBP algorithm is introduced. By analyzing the characteristics of the EBP algorithm to adjust the weight matrix, the idea of sample contribution degree is integrated into it, and the EBP algorithm based on sample contribution degree (SCBP) is obtained.

1) EBP ALGORITHM
The EBP algorithm has been the training foundation of many current neural network models since it was proposed by Rumelhart et al. [41] in 1986. The basic idea is that the learning process is caused by the forward propagation of signals and the back propagation of errors; two processes are therefore composed. Fig 1 illustrates a basic three-layer EBP neural network. The information of the entire network is stored in the weight matrix. When a sample value vector is input, the network calculates the actual output according to the weight matrix. Each sample has an expected output. The error between the expected and actual output is calculated. Based on this error, the error is reversely propagated, and the gradient descent algorithm is used to adjust the weight between the layers to achieve the purpose of memory.

2) SCBP ALGORITHM
The EBP algorithm is primarily adjusted based on the prediction error of a single input sample. The concept of the sample contribution was added to the adjustment strategy
TABLE 1. The explanation of symbols and variables.

| Symbols and variables | Explanation |
|-----------------------|-------------|
| q                     | The sample contribution of training sample |
| p                     | Number of training sample |
| x                     | Input value |
| y                     | Output of hidden layer |
| o                     | Actual output of the hidden layer |
| d                     | Expected output of the network |

The symbols are summarized in the Table 1.

a: DEFINITIONS OF SYMBOLS

The symbols are summarized in the Table 1.

Input layer:

The total virtual input of the p-th input sample is $\vec{x}'_p = (x'_p1, x'_p2, \ldots, x'_pm)$, and $x'_pi$ can be calculated by equation (1).

$$x'_pi = \sum_{t=1}^{p} (q_t \times x_{it})$$  (1)

Hidden layer:

The total virtual output of the hidden layer of the p-th input sample is $y'_pj = (y'pj1, y'pj2, \ldots, y'pm)$, and $y'pj$ can be calculated by equation (2) and (3).

$$y'pj = \sum_{i=1}^{p} (q_t \times y_{ij})$$  (2)

$$y_{ij} = f(net_{ij})$$  (3a)

$$net_{ij} = \sum_{i=1}^{n} (v_{ij} \times x_{ij})$$  (3b)

It is assumed that transfer functions $f(\cdot)$ are unipolar sigmoid functions, and $f'(x) = f(x) \times [1 - f(x)]$.

Output layer:

The output of the total virtual output of the p-th input sample is $o'_p(o'_p1, o'_p2, \ldots, o'_pm)$, and $o'_pk$ can be calculated by equation (4) and (5).

$$O'_p = \sum_{t=1}^{p} (q_t \times o_{tk})$$  (4)

$$o_{tk} = f(net_{tk})$$  (5a)

$$net_{tk} = \sum_{j=1}^{m} (w_{jk} \times y_{ij})$$  (5b)

$$d'_pk = \sum_{t=1}^{p} (q_t \times d_{tk})$$  (5c)

b: TRAINING PROCESS

The learning process is caused by the forward propagation of signals and the back propagation of errors. The forward process calculates the actual output of the network, and the reverse process adjusts the weight matrix. When the expectation does not match the output, the weighted output error is $E$, which is defined as equation (6).

$$E = \frac{1}{2} (D - O)^2 = \frac{1}{2} \sum_{k=1}^{l} \sum_{t=1}^{p} q_t \times (d_{tk} - o_{tk})^2$$  (6)

It is expanded to the hidden layer:

$$E = \frac{1}{2} \sum_{k=1}^{l} \sum_{t=1}^{p} q_t \times (d_{tk} - f(net_{tk}))^2$$  (7a)

$$E = \frac{1}{2} \sum_{k=1}^{l} \sum_{t=1}^{p} q_t \times (d_{tk} - f(\sum_{j=1}^{m} w_{jk} \times y_{ij}))^2$$.  (7b)

When further expanded to the input layer,

$$E = \frac{1}{2} \sum_{k=1}^{l} \sum_{t=1}^{p} q_t \times (d_{tk} - f(\sum_{j=1}^{m} \sum_{i=1}^{n} v_{ij} \times x_{ni}))^2.$$  (8)

From equation (8), it is evident that the error $E$ is a function of $w_{jk}$ and $v_{ij}$, and the adjustment rule is:

$$\Delta w_{jk} = -\eta \frac{\partial E}{\partial w_{jk}} \quad j = 1, 2, \ldots, m; \quad k = 1, 2, \ldots, l$$  (9a)

$$\Delta v_{ij} = -\eta \frac{\partial E}{\partial v_{ij}} \quad i = 1, 2, \ldots, n; \quad j = 1, 2, \ldots, m$$  (9b)

By applying equations (7-8), the following is obtained:

$$\Delta w_{jk} = -\eta \frac{\partial E}{\partial w_{jk}} = -\eta \sum_{t=1}^{p} q_t \times \left( \frac{\partial E}{\partial net_{tk}} \times \frac{\partial net_{tk}}{\partial w_{jk}} \right)$$  (10a)

$$\Delta v_{ij} = -\eta \frac{\partial E}{\partial v_{ij}} = -\eta \sum_{t=1}^{p} q_t \times \left( \frac{\partial E}{\partial net_{ij}} \times \frac{\partial net_{ij}}{\partial v_{ij}} \right).$$  (10b)

The error signals defined for the hidden layer and output layer are respectively:

$$\delta_{net}^{p} = -\frac{\partial E}{\partial net_{tk}}.$$  (11a)
Combined with equations (1-4) and (10-11), the following is obtained:

$$
\Delta w_{jk} = \eta \times y_p' \times \sum_{i=1}^{p} \delta_{ik}', \quad (12)
$$

$$
\Delta v_{ij} = \eta \times x_i' \times \sum_{t=1}^{p} \delta_{ij}'. \quad (13)
$$

The following is obtained by applying equations (6-7).

$$
\delta_{ik}' = -\frac{\partial E}{\partial net_{tk}} = (d_{ik} - o_{ik}) \times f'(net_{tk}) \quad (14)
$$

$$
\delta_{ij}' = -\frac{\partial E}{\partial net_{ij}} \times \frac{\partial y_j}{\partial net_{ij}} = \frac{\partial E}{\partial y_j} \times f'(net_{ij}) \quad (15)
$$

$$
\delta_{ik}' = \sum_{k=1}^{l} [(d_{ip}' - o_{ip}') \times f'((net_{tk}) \times w_{jk})] \times f'(net_{ij}) \quad (16)
$$

$$
\delta_{ik}' = (d_{ik} - o_{ik}) \times o_{ik}(1 - o_{ik}) \quad (17)
$$

$$
\delta_{ij}' = (\sum_{k=1}^{l} \delta_{ik}' \times w_{jk}) \times y_j \times (1 - y_j) \quad (18)
$$

Additionally, the following is defined:

$$
\delta_{ik}' = \sum_{t=1}^{p} \delta_{ik}'. \quad (19a)
$$

$$
\delta_{ij}' = \sum_{t=1}^{p} \delta_{ij}'. \quad (19b)
$$

By applying equations (1, 2, 12, and 13), the final result is:

$$
\Delta w_{jk} = \eta \times (d_{ip}' - o_{ip}') \times o_{ip}' \times (1 - o_{ip}') \times y_p', \quad (20)
$$

$$
\Delta v_{ij} = \eta \times (\sum_{k=1}^{l} \delta_{ik}' \times w_{jk}) \times y_p' \times (1 - y_p') \times x_i'. \quad (21)
$$

Compared with the traditional EBP algorithm, SCBP is no longer based on the prediction error of the current sample in each adjustment of the network weight, but comprehensively refers to all the samples trained, and uses the weighted prediction error as the adjustment basis. It enhances the memory and utilization of training samples, especially for training samples with obvious timing characteristics.

### C. GREY WOLF OPTIMIZER ALGORITHM

The GWO algorithm is a group intelligent optimization algorithm proposed by Mirjalili [42] in 2014. The algorithm was inspired by the prey activity of the grey wolf, and is an optimized search method. It is characterized by strong convergence performance, few parameters, and easy implementation. It has been of great concern for scholars in recent years, and has been successfully applied to the fields of shop scheduling [43], parameter optimization [44], and image classification [45], among others.

The grey wolf population in nature has a strict hierarchical system, and the wolf that is mainly responsible for making decisions on predation, habitat, and other activities is denoted as the $\alpha$ wolf.

The $\beta$ wolf assists the $\alpha$ wolf. When the $\alpha$ wolf is old, it becomes a candidate for the $\alpha$ wolf, and dominates the other ranks of wolves.

The $\delta$ wolves obey the $\alpha$ and $\beta$ wolves, and dominate the remaining wolf level. They are generally composed of young, sentinel, and hunting wolves.

The $\omega$ wolves must usually obey the wolves of other ranks.

The GWO optimization process includes the steps of the grey wolf’s social hierarchy, tracking, enveloping, and attacking prey. The mathematical model for simulating the swarm behavior is as follows.

$$
D = C \times X_p(t) - X(t) \quad (22)
$$

$$
X(t + 1) = X_p(t) - A \times D \quad (23)
$$

$$
A = 2\gamma_1 - \alpha \quad (24)
$$

$$
C = 2\gamma_2 \quad (25)
$$

$$
D_\alpha = C_1 \times X_\alpha - X, \quad D_\beta = C_2 \times X_\beta - X, \quad D_\delta = C_3 \times X_\delta - X \quad (26)
$$

$$
X_1 = X_\alpha - A_1 \times D_\alpha, \quad X_2 = X_\beta - A_2 \times D_\beta, \quad X_3 = X_\delta - A_3 \times D_\delta \quad (27)
$$

$$
X(t + 1) = \frac{X_1 + X_2 + X_3}{3} \quad (28)
$$

In equations, (22-28), $t$ is the current iteration, and $A$ and $C$ are coefficients of concordance, $X_p(t)$ is the prey position, $X(t)$ is the location of each wolf in the population, $\alpha$ linearly decreases from 2 to 0 as the number of iterations increases, and $\gamma_1$ and $\gamma_2$ are random vectors from the interval $(0, 1)$. The pseudocode of GWO is as follows.

#### D. GWO-SCBP ALGORITHM

By using the optimization strategy of the GWO algorithm, the optimization of the sample contribution is considered. To obtain the optimal combination of each sample contribution, the EBP algorithm based on the optimal sample contribution algorithm (GWO-SCBP) is finally established. The steps are presented in Fig. 2, and the specific steps are as follows.

Step 1: Input the training sample and random sample contributions;

Step 2: Use the SCBP algorithm to obtain the trained model and predictive value;

Step 3: Use the predictive value and actual value to obtain the fitness;

Step 4: Sort the population according to the degree of fitness, and obtain the wolf class;

Step 5: The population searches for prey according to the management strategy of the $\alpha$, $\beta$, and $\gamma$ wolves.

Step 6: Update the positions of the wolves and limit the group position;
Algorithm 1  Pseudocode of Grey Wolf Optimizer Algorithm

**Input:** group number, maximum iterations, position limit, individual dimension;
**Output:** history fitness, $\alpha$ wolf fitness, $\alpha$ wolf position;

1. Generating initial population;
2. while $i \leq$ maximum iterations do
   3. // Electing all classes of wolves;
   4. for each wolf do
      5. fitness = predictive value - actual value;
      6. end
   7. First fitness is $\alpha$ wolf, second is $\beta$ wolf, third is $\delta$ wolf, others are $\omega$ wolf;
   8. // Update the position of each class of wolf;
   9. Calculate each parameter based on Eq. (22) to Eq. (25)
   10. $\alpha$, $\beta$, $\delta$ wolves guide candidate wolf to close to the prey, and the positions of all wolves are updated based on Eq. (26) to Eq. (28)
   11. // Limit the position of the wolf group;
   12. $i = i + 1$;
13. end
14. return history fitness, $\alpha$ wolf fitness, $\alpha$ wolf position.

**FIGURE 2.** GWO-SCBP computing framework.

**A. STUDY AREA**

Xi’an is the capital city of Shaanxi Province and is an important cultural, economic, and educational city in China. By the middle of 2018, Xi’an had opened 3 subway lines, namely Line 1, Line 2, and Line 3. There are 63 stations, including 3 interchange stations.

The first metro line, Line 2, began operation on September 16, 2011. Xi’an Metro Line 2 runs along the central axis of the city, connecting Xi’an’s administrative center, transportation hub, university gathering area, scientific research area, commercial center, and many tourist attractions.

Xi’an Metro Line 1 began operation on September 15, 2013. It was the first rail transit line to connect Xi’an and the Xianyang New Area. The beginning of the operation of the second phase of Xi’an Metro Line 1 is of great significance for accelerating the integration of Xi’an and Xianyang, optimizing the urban layout structure, and promoting the rapid development of the urban social economy.

The last metro line, Line 3, began operation on November 8, 2016. It connects the Xi’an International Port Area and the Xi’an High-tech Industrial Development Zone, which is the skeleton line of Xi’an rail transit network planning. Therefore, it is necessary to establish an efficient productive system of metro passenger flow. For the convenience of research, the labels for each station are listed in Fig. 3.

**B. DATA ACQUISITION**

Due to the popularity of digitalization, the use of automatic fare collection (AFC) systems offers the possibility of the collection of urban rail transit data. Therefore, the out-flow from May 2017 to August 2018 was collected from this dataset. There were more than 560,000 statistical data points, and, after processing, 28,000 daily out-flow data points were calculated.

**C. OPTIMIZATION FRAMEWORK**

Due to the large amount of data selected, it was unrealistic to determine the contribution of a single sample in the prediction process. Therefore, station No. 41 (YHZ) was taken as an example, and its out-flow within one year is presented in Fig. 4. Its weekly out-flow exhibited the same fluctuation trends in different months and had the same cyclical variation in the same month. In other words, the distributions of daily sample contributions in the same week of different months were the same. Therefore, a hierarchical daily contribution calculation method was established. The steps were as follows.

Step 1: The training sample data was divided into 12 months to determine the contribution of each sample month to the forecast month;

Step 2: According to the weekly classification, the weekly contribution of different days in a week to the forecast month was examined from Monday to Sunday;
Step 3: Combined with the monthly sample contribution and the weekly sample contribution, the contributions of the daily samples were obtained.

The mathematical model is as follows.

It is assumed that the sample monthly out-flow contribution for the forecasted monthly out-flow is $C_m$, which is defined as:

$$ C_m = (\tilde{c}_{m1}, \tilde{c}_{m2}, \ldots, \tilde{c}_{mi}) \quad i = 1, 2 \ldots 63. \quad (29) $$

The $\tilde{c}_{mi}$ vector is:

$$ \tilde{c}_{mi} = (c_{mi1}, c_{mi2}, \ldots c_{mi2}) \quad j = 1, 2 \ldots 12, \quad (30) $$

where $i$ is the number of stations and $j$ is the sample month number. The sample weekly out-flow contribution for the forecasted weekly out-flow is $C_w$, which is defined as:

$$ C_w = (c_{w1}, c_{w2}, \ldots, c_{wi}). \quad (31) $$

The matrix of $c_{wi}$ is:

$$ c_{wi} = (\tilde{c}_{wi1}, \tilde{c}_{wi2}, \ldots, \tilde{c}_{wik})^T \quad k = 1, 2 \ldots 7, \quad (32) $$

and the $\tilde{c}_{wik}$ vector is:

$$ \tilde{c}_{wik} = (c_{wi1k}, c_{wi2k}, \ldots, c_{wikk}). \quad (33) $$

where $k$ is the number of days in a week.

The daily contribution of each sample $C_r(c_{r1}, c_{r2}, \ldots, c_{rt})$ is then obtained by the monthly sample contribution and weekly sample contribution. The matrix of $c_{ri}$ is:

$$ c_{ri} = \begin{bmatrix} c_{r1,1} & \cdots & c_{r1,t} \\ \vdots & \ddots & \vdots \\ c_{r1,k} & \cdots & c_{r1,t} \end{bmatrix} \quad t = 1, 2 \ldots 84, \quad (34) $$

where the $t$ is the number of all days in one week in all sample months. $c_{ri}$ is calculated by equation (35).

$$ c_{ri} = c_{wi} \times \tilde{c}_{mi} \quad (35) $$

The specific framework is illustrated in Fig. 5. The hierarchical calculation method can effectively reduce the dimension of the optimal contribution group, making it possible to find the optimal contribution. Table 2 compares the hierarchical calculation method (HCM) with the single sample method (SSM).
FIGURE 5. Hierarchical contribution calculation framework. The daily sample contribution was obtained from the weekly and monthly sample contributions.

TABLE 2. Comparison of dimensions of the SSM and HCM.

| Method | SSM | HCM |
|--------|-----|-----|
| Dimension | 365 | 12/7 |

D. ALGORITHM COMPARISON AND PARAMETER SETTINGS

To validate the optimization ability of GWO, particle swarm optimization (PSO) was introduced as the benchmark for comparison. The experimental environment included MATLAB 2018b, self-written MATLAB programs, and a computer with an Intel(R) Core (TM)-i5 and Windows 10 operating system.

The parameters of the PSO algorithm were set as follows: the individual and group learning factor was 2.0, the inertia weight was 0.4 and 0.8. The number of iterations and the particle swarm number was 50. The parameters of the GWO algorithm were set as follows: the number of iterations and the wolf group number was 50.

E. CALCULATION PROCESS

Taking the contribution of out-flow at station No. 41 as an example, the value is $c_{w41} \times c_{m41}$. In the subsequent section, a detailed description of the determination of the sample contribution for station No. 41 is provided.

1) MONTHLY SAMPLE CONTRIBUTION

Based on the monthly out-flow from May 2017 to April 2018, the total out-flow in May 2018 was predicted on a monthly basis. The fitness of the EBP network is defined as follows:

$$fitness_m = \text{abs}(\text{mean}(H_n) - c_n), \hspace{1cm} (36)$$

where $H_n$ is the predictive value, $c_n$ is the actual value(month), $n$ is set to eliminate random fluctuations in network prediction, and $n = 10$. The monthly sample contribution and the iterative process of the sample obtained by the GWO-SCBP algorithm and PSO-SCBP algorithm are presented in Fig. 6. It proves that GWO has a good effect.

2) WEEKLY SAMPLE CONTRIBUTION

The out-flow of all Mondays in May 2018 was predicted on a weekly basis. The fitness of the EBP network is defined as follows:

$$fitness_w = \text{sum}[\text{abs}(\text{mean}(H_{n4}) - c_{n4})], \hspace{1cm} (37)$$

where $H_{n4}$ is the predictive value (the forecast month has four Mondays), $c_{n4}$ is the actual value(four Mondays), $n$ is set to eliminate random fluctuations in network prediction, and $n = 5$. The weekly sample contribution and the iterative process of the sample obtained by the GWO-SCBP algorithm and PSO-SCBP algorithm are presented in Fig. 7.
W. Lu et al.: Research on Sample Selection

Fig. 8. Daily sample contribution of the prediction Monday of station No. 41 (YHZ) (only one week per sample month).

### TABLE 3. Definition of all prediction method.

| Method   | Submethod      | Feature                      |
|----------|----------------|-------------------------------|
| BPNN     |                | Without considering contribution |
| Ag-BPNN  |                | Average sample contribution   |
| G-BPNN   |                | Optimal sample contribution   |
| Elman    |                | Without considering contribution |
| Ag-Elman |                | Average sample contribution   |
| G-Elman  |                | Optimal sample contribution   |
| LSTM     |                | Without considering contribution |
| Ag-LSTM  |                | Average sample contribution   |
| G-LSTM   |                | Optimal sample contribution   |

### V. RESULTS AND ANALYSIS

In this section, the experimental result is verified. In addition, the factors that affect the sample contribution in aspect of the training sample features and the prediction sample features are analyzed.

#### A. RESULTS VERIFICATION

The contribution of the daily out-flow of urban rail transit depends largely on the contribution of monthly out-flow, i.e., the monthly sample size is used to verify the results of the sought sample contribution. The BPNN, Elman, and LSTM neural network based on the EBP algorithm were selected to verify the optimization results. The average contribution, optimal contribution, and neural network without considering contribution were also examined for comparison. The methods are defined in Table 3.

1) **ELMAN NEURAL NETWORK**

The Elman neural network, also known as the SRN (simple recurrent network), and was proposed by Elman in 1990. The SRN considers the timing information, and the output is not only related to the input of the current time, but also the input of all the previous moments. The SRN is the simplest of the RNN architectures. It adds timing feedback connections only to the fully connected layer as compared to traditional two-layer fully connected feedforward networks.

2) **LSTM NEURAL NETWORK**

As a variant of the Recurrent Neural Network (RNN), LSTM is used to selectively forget the useless information in the learning process of neural network. It includes forget gate, input gate, and output gate. The forget gate can decide which information should be retained and discarded. The information from the previous hidden state and the current input are input into the Sigmoid function at the same time. The output value is between 0 and 1. The closer to 0 means the more you should forget. The input gate is used to update the unit status. First input the information of the previously hidden state and the current input information into the Sigmoid function, and adjust the output value between 0 and 1 to determine which information to update. The output gate determines the value of the next hidden state. The hidden state contains information about the previous input.

3) **VALIDATION RESULTS**

These methods were used to predict the out-flow of station No. 41 (YHZ) in May 2018 to verify the optimization results. To evaluate and compare the forecasting performance of the proposed models and other alternatives, three common forecasting errors were selected as indices of performance assessment, namely the root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE). The RMSE, MAE and MAPE can be calculated as follows:

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{10} (y_i - \hat{y})^2},
\]

\[
MAE = \frac{1}{n} \sum_{i=1}^{10} |y_i - \hat{y}|
\]

\[
MAPE = \frac{1}{n} \sum_{i=1}^{10} \left( \frac{|y_i - \hat{y}|}{\hat{y}} \right) \times 100\%
\]

where \(y_i\) is the forecasted value, \(\hat{y}\) is the average of forecasted value, \(\hat{y}\) is the actual value and \(n\) is the number of predictions. The prediction errors are shown in Fig. 9 and Fig. 10.

It is evident from Fig.9 and Fig.10 that the search for the optimal sample contribution achieved good results in these prediction methods, indicating that the optimization result is credible.
B. RESULT ANALYSIS

Studies have shown that the impact of training and prediction sample selection for the model. The factors that affect the sample contribution in aspect of the training sample features and the prediction sample features are analyzed.

1) TRAINING SAMPLE

Due to the characteristics of the EBP algorithm, the predicted value is determined by the input and output of the training sample. The distribution of the monthly sample contribution of the search in both the temporal and spatial dimensions is analyzed.

a: TEMPORAL DIMENSION

In the temporal dimension, the Euclidean distance between the sample and predicted input, the absolute distance between the sample month and forecast month out-flow, and the Pearson coefficient of the sample contribution with the distance were calculated. The Pearson coefficients of each station were calculated separately, and the results are presented in Fig. 11.

It can be seen from Fig. 11 that the Pearson coefficients of most stations were in the range of $[-0.5, 0.5]$, indicating the weak correlation between input/output variables distance with the sample contribution. The relationship between input/output distance and sample contribution for stations No. 5 (HCL) and No. 14/54 (THM) is presented in Fig. 12.

From the figure, it can be found that the larger the distance between the sample and predicted input, the higher the contribution of the sample. The larger the distance between the sample and predicted output, the smaller the contribution of the sample.

b: SPATIAL DIMENSION

In the spatial dimension, the sample contribution distributions of different stations are different, so it is necessary to analyze the relationship between the sample contribution distribution and the station characteristics. The Pearson coefficient of the monthly out-flow and the sample contribution distribution at each station were calculated, and the results are presented in Fig. 13.

It can be seen from Fig. 13 that except for the smaller absolute value Pearson coefficients of stations No. 14/54 (THM) and No. 20 (BKZ), the Pearson coefficients of all other stations were in the range of $[0.5, 1]$, indicating that the monthly out-flow at each station has a strong correlation with the monthly sample contributions.

Further, taking station No. 41 (YHZ) as an example, the relationship between monthly out-flow and sample contribution is presented in Fig. 14.

It can be found that the smaller the monthly out-flow, the greater the sample contribution of this month. In other words, the out-flow is more sensitive for months with small out-flow, and the more important it is for the forecast month, the greater its sample contribution.

2) PREDICT SAMPLE

To verify the adaptability of the sample contribution to different prediction months, the sample contributions were
FIGURE 14. Trend in monthly out-flow and sample contribution of station No. 41(YHZ).

FIGURE 15. Comparison of sample contribution of different prediction month. (The prediction months are from May 2018 to August 2018).

compared. Taking station No. 41 (YHZ) as an example, and the results are presented in Fig. 15.

It can be seen from Fig. 15 that there was the same trend of the sample contributions for different prediction months. In other words, the sample contribution trend does not change with the prediction months.

VI. DISCUSSION
In the previous section, the qualitative relationship between the out-flow, the input/output distance, and its monthly contribution distribution was given. In this section, stations are classified according to the relationship between the temporal and spatial dimensions and the distribution of the sample contributions. Based on the out-flow distance at each station, the central station of each type of station is calculated, and the sample contribution of each type of the central station is fitted.

A. STATION CLASSIFICATION
According to the results in Section 5, it was found that the influencing factors of sample contribution for each station are very different. The stations were thus divided into four categories based on the different influencing factors, and are defined as follows.

Type I: The sample contribution is related to the distance between the sample and the predicted input. It includes station No. 14/54 (THM).

Type II: The sample contribution is related to the distance between the sample and the predicted output. It includes station No. 20 (BKZ).

Type III: The sample contribution is not only related to the monthly out-flow trend, but also to the distance between the sample and the predicted output. It includes stations No. 5 (HCL), No. 10/29 (BDJ), No. 11 (WLK), and No. 34/47 (XZ).

Type IV: The sample contribution is related to the monthly out-flow trend. It includes all other stations.

B. CENTRAL STATION
In practical research, choosing a representative example is a common method to simplify the calculation for complex problems. Similarly, to effectively explain the characteristics of each type of station, it is necessary to select the central station of each type of station for further research. By measuring the distance, the central station of each type of station was determined.

1) FRECHET DISTANCE ALGORITHM
The concept of distance space was proposed by the French mathematician Frechet in 1906, and is also known as the Frechet distance. It extends the concept of distance in real space to a general set and provides a theoretical basis for distance measurement between abstract spaces. It focuses on taking the path-space distance into account, thereby improving the efficiency of the evaluation of the similarity of curves with a certain spatial timing.

A and B are two continuous curves on S, namely A:[0,1]→S and B:[0,1]→S. Additionally, α and β are two reparameterization functions of the unit interval, namely α:[0,1]→[0,1], β:[0,1]→[0,1]. The Frechet distance of curves A and B are therefore defined as:

\[ F(A, B) = \inf_{\alpha, \beta \in [0, 1]} \max_{t \in [0, 1]} d(\alpha(\alpha(t)), B(\beta(t))) \]  \hspace{1cm} (41)

where d is the metric function on S. The smaller the Frechet distance, the higher the similarity between the two curves.

2) CENTRALIZATION PROCESS
The central station of each type of station can be calculated by measuring the distance, and the specific steps are as follows.

Step 1: For each station, the Min-Max Normalization theory is used to normalize historical month out-flow;

Step 2: Calculate the Frechet distance between the normalized out-flow values of any two stations of the same type of station;

Step 3: The station with the smallest sum of distances from other stations is the central station of this type of station.

The results are presented in Fig. 16 and Table 4.

| TABLE 4. Central station calculation results. |
|-----------------------------------------------|
| Station class | Type I | Type II | Type III | Type IV |
| Central station | No. 14/54 | No. 20 | No. 34/47 | No. 6 |

VOLUME 8, 2020
where $x$ and $R$.

Types I and II each contain only one station.

FIGURE 17. Various center station fitting results. The Central station calculation results (Type III and Type IV).

cause by the time series. The fitting equation is:

$$f(x) = (4.895e+04) * x - 0.049,$$

where $x$ is the distance between the sample and forecast input, and $R^2 = 0.95$.

Type III stations include four stations, namely stations No. 5 (HCL), No. 10/29 (BDJ), No.11 (WLK), and No.34/47 (XZ). Station No.10/29 (BDJ) is the interchange station for Line 1 and Line 2, and station No.34/47 (XZ) is the interchange station for Line 2 and Line 3. They are located on the axis of urban development, and there are many surrounding businesses. Stations No.5 (HCL) and No.11 (WLK) are passenger stations. These four stations undertake seasonal passenger and commuter flow. Their monthly sample contributions are affected by regular changes of seasonal passenger flow and commuter flow. The central station fitting equation is:

$$f(x_1, x_2) = (-6.893e-07) * x_1 + (-2.310e-07) * x_2,$$

where $x_1$ is the distance between the sample and forecast output, $x_2$ is the monthly out-flow, and $R^2 = 0.65$.

Type IV stations include all other 57 stations. Their monthly sample contributions are affected by commuter passenger flow. The central station fitting equation is:

$$f(x) = (3.855e+05) * x - 0.901,$$

where $x$ is the monthly out-flow countdown, and $R^2 = 0.81$.

C. TYPE OF STATION ANALYSIS

As determined by the temporal dimension, the first to third types of stations are generally interchange stations or tourist stations, and their out-flow is greatly affected by the fluctuation of the time series. The fourth type of station is generally a commuter station, and due to the elimination of the disturbance of major holidays, its out-flow changes are less affected by the time series, and its random fluctuations occupy a significant position on the impact of such stations. Further, specific expressions are given to various central stations via linear fitting in Fig. 17.

Type I station includes only one station, No. 14/54 (THM), which is the interchange station for Line 1 and Line 3. Its function is primarily to undertake the interchange passenger flow, and its passenger flow is small and stable. Its monthly sample contribution is affected by sudden passenger flow caused by the time series. The fitting equation is:

$$f(x) = 0.0469 * x - 0.057,$$

where $x$ is the distance between the sample and forecast input, and $R^2 = 0.67$.

FIGURE 16. Central station calculation results (Type III and Type IV). Types I and II each contain only one station.

W. Lu et al.: Research on Sample Selection

VII. CONCLUSION

In this paper, the concept of sample contribution was proposed and integrated into the classical EBP algorithm. The EBP algorithm with the consideration of sample contribution was deduced, and the SCBP algorithm was obtained. The GWO algorithm was used to optimize the sample contribution degree, and the optimization framework of the sample contribution degree was constructed (GWO-SCBP). Taking the out-flow forecast of each station of the Xi’an urban rail network as an example, the contribution of historical samples in the forecasting process was optimized. The multiple prediction method was used to verify the optimization results, which were demonstrated to be credible. The optimization framework is therefore feasible.

Furthermore, the optimization results were analyzed in aspect of the training sample features and the prediction sample features. It was found that the distribution of the historical sample contribution of each station is not only related to the time series, but also has a strong connection with the unique out-flow changes of each type of station.

Finally, by selecting various types of central sites, specific fitting formulas were provided, and were found to obtain good fitting effects.
Additionally, for urban rail transit passenger flow forecasting, although months with large contributions are more important, months with lesser contributions must also be included to create a complete time series to accurately reflect the characteristics of historical passenger flow. Although the SCBP algorithm was used to study the sample contribution in urban rail passenger flow prediction, holidays, weather, events, and sudden passenger flow have always been a major difficulty in prediction. Therefore, research on different types of time series and passenger flow should be carried out in future work.

**REFERENCES**

[1] Y. Wang and C. Ma, “Influencing factors and development trends of urban rail transit passenger flow,” J. Chang’an Univ. (Natural Sci. Ed.), vol. 33, no. 3, pp. 69–75, 2013.

[2] Y. X. Xu, X. Zhai, Y. C. Liu, and Q. J. Kong, “Short-term prediction method of freeway traffic flow,” J. Traffic Transp. Eng., vol. 13, no. 2, pp. 114–119, 2013.

[3] M. Milenković, L. Švadlenka, V. Melichar, N. Bojić, and Z. Avramović, “SARIMA modelling approach for railway passenger flow forecasting,” Transport, vol. 33, no. 5, pp. 1113–1120, 2018, doi: 10.3846/16484142.2016.1396263.

[4] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik, “Support vector regression machines,” in Proc. Adv. Neural Inf. Process. Syst., vol. 9, M. C. Mozer, M. I. Jordan, and T. Petsche, Eds. Cambridge, MA, USA: MIT Press, 1997, pp. 155–161.

[5] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Comput., vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[6] Z. Zhao, W. Chen, X. Wu, P. C. Y. Chen, and J. Liu, “LSTM network: A deep learning approach for short-term traffic forecast,” IET Intell. Transp. Syst., vol. 11, no. 2, pp. 68–75, Mar. 2017.

[7] 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, no. 1, pp. 1–9, Feb. 2019, doi: 10.1155/2019/8392592.

[8] L. Qin, W. Li, and S. Li, “Effective passenger flow forecasting using STL and ESN based on two improvement strategies,” Neurocomputing, vol. 356, pp. 244–256, Sep. 2019, doi: 10.1016/j.neucom.2019.04.061.

[9] M. Ni, Q. He, and J. Gao, “Forecasting the subway passenger flow under event occurrences with social media,” IEEE Trans. Intell. Transp. Syst., vol. 18, no. 6, pp. 1623–1632, Jun. 2017.

[10] H. Wang, L. Li, P. Pan, Y. Wang, and Y. Jin, “Early warning of burst passenger flow in public transportation system,” Transp. Res. Part C, Emerg. Technol., vol. 105, pp. 580–598, Aug. 2019.

[11] G. M. Foody, M. B. McCulloch, and W. B. Yates, “The effect of training set size and composition on artificial neural network classification,” Int. J. Remote Sens., vol. 16, no. 9, pp. 1707–1723, Jun. 1995.

[12] C. Li, J. Wang, L. Wang, L. Hu, and P. Gong, “Comparison of classification algorithms and training sample sizes in urban land classification with landset thematic mapper imagery,” Remote Sens., vol. 6, no. 2, pp. 964–983, 2014.

[13] Z. Zhen, L. J. Quackenbush, S. V. Stehman, and L. Zhang, “Impact of training and validation sample selection on classification accuracy and accuracy assessment when using reference polygons in object-based classification,” Int. J. Remote Sens., vol. 34, no. 19, pp. 6914–6930, 2013, doi: 10.1080/01431161.2013.810822.

[14] M. B. Lyons, D. A. Keith, S. R. Phinn, T. J. Mason, and J. Elith, “A comparison of resampling methods for remote sensing classification and accuracy assessment,” Remote Sens. Environ., vol. 208, pp. 145–153, Apr. 2018.

[15] B. Kalantar, B. Pradhan, S. A. Naghibi, A. Motevalli, and S. Mansor, “Assessment of the effects of training data selection on the landslide susceptibility mapping: A comparison between support vector machine (SVM), logistic regression (LR) and artificial neural networks (ANN),” Geomatics Natural Hazards Risk, vol. 9, no. 1, pp. 49–69, 2018.

[16] R. Setiono, A. Azzarraga, and Y. Hayashi, “Using sample selection to improve accuracy and simplicity of rules extracted from neural networks for credit scoring applications,” Int. J. Comput. Intell. Appl., vol. 14, no. 4, Dec. 2015, Art. no. 1500021.
[43] X.-Y. Lv and T. L. Liao, “Permutation flow-shop scheduling based on the grey wolf optimizer,” J. Wuhan Univ. Technol., vol. 37, no. 5, pp. 111–116, 2015.

[44] B. Rabhi, H. Dhahri, A. M. Alimi, and F. A. Alturki, “Grey wolf optimizer for training elman neural network,” in Proc. 16th Int. Conf. Hybrid Intell. Syst., vol. 552, A. Abraham, A. Haqiq, A. M. Alimi, G. Mezzour, N. Rokbani, and A. K. Mada Eds. Cham, Switzerland: Springer, 2017, pp. 380–390.

[45] S. Kapoor, I. Zeya, C. Singhal, and S. J. Nanda, “A grey wolf optimizer based automatic clustering algorithm for satellite image segmentation,” in Proc. 7th Int. Conf. Adv. Comput. Commun., vol. 115, M. Gahegan and J. P. Mulerikkal, Eds. Amsterdam, The Netherlands: Elsevier 2017, pp. 415–422.

**WENBO LU** received the B.S. degree in traffic engineering from the Xi’an University of Architecture and Technology, in 2018. He is currently pursuing the M.S. degree from Chang’an University, China. His research interests include transportation planning and management, deep learning, and the passenger flow forecasting of urban rail transit.

**PEIKUN LI** received the B.S. degree in light chemical engineering from Northeast Electric Power University, Jilin, China, in 2015. He is currently pursuing the master’s degree with the Chang’an University, China. His current research interests include traffic big data analysis, commuter behavior analysis, and deep learning transportation applications.

**CHAOQUN MA** received the B.S., M.S., and Ph.D. degrees in transportation planning and management from Chang’an University, China, in 2000, 2003, and 2007, respectively. He is currently an Associate Professor with the College of Transportation Engineering, Chang’an University. His current research interests include traffic behavior analysis and rail transit planning and passenger flow prediction.

**VOLUME 8, 2020**