Big Data Multi-Strategy Predator Algorithm for Passenger Flow Prediction

Peng Guo
School of Management, Zhengzhou University of Industrial Technology, Zhengzhou, 450064, China

***Abstract*—Faced with the rapidly recovering tourism market, accurate prediction of passenger flow can help local authorities achieve more effective resource regulation. Therefore, based on big data technology, a multi-strategy predator algorithm is proposed, which uses the Marine Predator Algorithm, combined with regularized extreme learning machines and Collaborative Filtering Algorithms, to achieve accurate passenger flow prediction. The experiment findings denote that the performance parameters of the algorithm are excellent, with extremely strong convergence performance, and only 30 iterations are needed to reach the optimal solution. The fitting degree of this algorithm is 97.8%, which is 6.27% -19.31% higher than that of long and short-term memory networks, random forest algorithms, and support vector machine regression. In actual passenger flow prediction, the error rate of this algorithm is only 2.29%, which is 3.47% - 6.50% higher than the three comparison algorithms. This study provides a new and efficient prediction method for passenger flow prediction. Its excellent predictive performance can not only help relevant departments predict and manage passenger traffic more accurately, but also provide reference for traffic prediction in other fields. Overall, this study has important reference value and practical significance for the research and practice of passenger flow prediction.**

Keywords*—Passemger Flow Prediction; regularized extreme learning machine; Collaborative Filtering Algorithm; Marine Predator Algorithm

I. Introduction

In today's globalized world, Passenger Flow Prediction (PFP) has become a key factor in determining the competitiveness of industries such as tourism, transportation, and hotels [1]. Accurate and rapid predictions can help local tourism departments make strategic arrangements in advance and maximize resource utilization efficiency [2]. However, accurate PFP is extremely difficult, as passenger flow is influenced by many complex factors, including seasonality, weather conditions, holidays, policy adjustments, etc., resulting in highly uncertain prediction results [3-4]. Big data technology enables researchers to collect large amounts of passenger flow data from multiple sources, including but not limited to social media, online travel platforms, traffic monitoring systems, and more. Big data technology supports the operation of complex machine learning algorithms, such as Regularized Extreme Learning Machine (RELM) and Collaborative Filtering Algorithm (CFA), which are trained on large-scale data sets and can improve the accuracy and generalization ability of prediction models [5]. The study combines RELM, Marine Predator Algorithm (MPA), and CFA to maximize prediction accuracy, and uses time series reconstruction to process temporal correlation information. The innovation of the research is reflected in two aspects. Firstly, the use of big data technology enables the model to utilize more information and improve the accuracy of predictions. The second is the combination of three algorithms, which can better handle the complexity and uncertainty in PFP. Research can not only improve the accuracy of PFP, thereby helping enterprises make better strategic decisions, but also provide new ideas and methods for subsequent PFP research. In addition, the application of this method is not limited to PFP, but can also be used in other fields that require prediction, such as electricity demand prediction, stock market prediction, etc., with broad application prospects. The contributions of the research are as follows: (1) a multi-strategy predator algorithm based on big data technology is proposed, which integrates RELM, MPA and CFA, effectively improving the accuracy of PFP. (2) The research shows the practical application of big data technology in PFP. By processing and analyzing a large number of passenger flow data, the training quality of the forecasting model and the accuracy of the forecasting results are improved. (3) The research not only provides a new and efficient method for the field of PFP, but also provides reference and enlightenment for the application of big data technology in other related fields. The study is composed of four parts. The first part is a brief explanation of the relevant field of research, the second part is the implementation of the proposed method, the third part is the testing and validation of the proposed method, and the fourth part is a summary and outlook of the research content.

II. Related Works

PFP is a technique used to estimate the number of people in a specific area or track the direction of human flow. This technology is widely used in multiple fields such as public safety, retail, transportation planning and management. For example, shopping malls may use PFP to understand customer shopping habits and flow paths, to better layout stores and products. The transportation department may use PFP to evaluate the condition of the transportation network for more effective traffic planning and management. Sevtsuk proposed a method for estimating pedestrian travel generation and distribution in urban streets for PFP. The research results were validated in the Kendall Square area of Cambridge, and the PFP was highly accurate compared to the observed number of passengers [6]. Cooper et al. proposed a regression direct demand model based on multiple mixed spatial design network analysis for PFP in complex urban environmental layouts. The experimental results showed that the model successfully predicted pedestrian flow of approximately 8000 people per hour [7]. Togashi et al. used a method combining Kalman

filtering for PFP, and the research outcomes indicated the practical value of Kalman filtering in PFP [8]. Zhang et al. raised a deep learning architecture that combines residual networks, graph convolutional networks, and long and short-term memory (LSTM) for short-term PFP in urban rail transit operations. The experiment findings demonstrated the advantages and robustness of this method [9]. Quan et al. put forward a PFP method based on LSTM, and the experimental results proved the performance of this method in road traffic safety [10].

The predator model is based on some basic biological assumptions, such as the growth rate of prey being proportional to its own amount, and the growth rate of predators being proportional to the amount of prey they prey on. In the fields of computational science and optimization algorithms, predator models are often used to solve global optimization problems, simulating the interaction between predators and prey in nature, aiming to find the global optimal solution of optimization problems through this simulation. Ramezani et al. proposed an improved version based on adversarial learning methods, chaotic graphs, population adaptation, and exploration and utilization stage switching to address the shortcomings of MPAs. Experimental results showed that this method has better performance [11]. Ahn et al. demonstrated the global existence and uniform boundedness of solutions for the general functional response model in any spatial dimension for the predator-prey model of indirect prey chemotaxis. Further linear stability analysis revealed that prey chemotaxis is a key factor in the formation of patterns, which is beneficial for promoting the further application of predator algorithms [12]. Ghanbari et al. considered an approximation of predator-prey interactions in the presence of prey social behavior for the time fractional derivative in a three species predator-prey model [13]. He et al. proposed a stochastic predator-prey model to address the problem of species extinction caused by environmental pollution. They established sufficient conditions for the average persistence and extinction of species. The analysis results were validated through numerical examples [14]. Bortuli et al. proposed a prey predator interaction mathematical model that divides prey populations into susceptible and infected categories to address the issue of predator selection of susceptible prey leading to population extinction. The different biological characteristics of the model were determined through numerical simulation and the analysis results were verified [15].

The main methods in the field of PFP are listed in detail, including the LSTM-based prediction method, the regression direct demand model based on multiple hybrid space design network analysis, and the prediction method combined with Kalman filter. Each of these methods has its advantages, but there are also limitations, such as the computational efficiency of LSTM in processing large-scale data, and the limitations of Kalman filtering in nonlinear problems. By comparing it with existing work, the study identifies areas where further work is needed, such as the optimization of the algorithm in handling data fluctuations and real-time predictions in extreme cases, as well as the improvement of generalization ability in different scenarios. This study not only analyzes the differences between the proposed method and existing methods in theory, but also makes a direct comparison in experiment. To fill the gap between the existing work and this study, the study proposes targeted strategies, including further optimization of the algorithm parameters, improving the algorithm's performance on non-linear and high-dimensional data, and exploring the algorithm's application potential in other fields. By comparing the performance of LSTM, Random Forest (RF), and Support Vector Regression (SVR), the advantages of the proposed algorithm in many performance indexes such as fitting degree, convergence speed and error rate are proved. Future research can consider further optimizing algorithms to improve computational speed. In addition, the results of this study are not limited to PFP, but can also be extended to other related fields, such as traffic flow prediction, market demand prediction, etc., demonstrating a wide range of application prospects.

III. Construction of Big Data Algorithms for Passenger Flow Prediction

The construction of big data MSP algorithms involves three core parts: the construction of multi-strategy algorithms, the construction of MPAs, and the optimization design of MSP algorithms. In the construction of multi-strategy algorithms, RELMs and CFAs are combined to effectively solve complex problems. In the construction of the MPA, the predatory and reproductive behaviors of marine predators are mainly simulated to achieve global search and local fine search of the problem solution space. In the optimization design of the MSP algorithm, time series reconstruction is used to process time data, and the algorithm is optimized to raise the search efficiency and quality of the solution.

A. Construction of RELM-CFA Multi-strategy Algorithm Model

To predict and handle the complexity and uncertainty of passenger flow more accurately, this study chooses to use RELM and CFA as a combination strategy. RELM is a single hidden layer feedforward neural network that introduces regularization terms on the basis of RELM, which can effectively handle overfitting problems of data and improve prediction accuracy [16]. Fig. 1 shows a schematic diagram of the framework of the RELM.

Assuming the activation function of ELM is $g(\cdot)$, the RELM model can be represented as shown in Formula (1).

$$\sum_{i=1}^{L} \beta_i g\left(\omega_i X_j + b_i\right) = Y_j,\ j = 1, 2, \cdots, N. \quad (1)$$

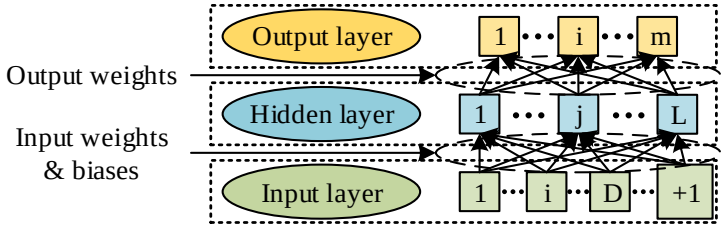

Fig. 1. Schematic diagram of the framework of the regularized extreme learning machine.

In Formula (1), ω_i represents the weight vector between the input layer and the i th hidden layer node. X_j represents the input feature of the j th sample. β_i represents the output layer weight between the i th hidden layer node and the predicted target. Y_j represents the predicted value of the j th sample. $\omega_i X_j$ represent the inner product of ω_i and X_j. b represents bias. Formula (1) is represented in matrix form, as shown in Formula (2).

$$\begin{cases} H\beta = Y \\ H = \begin{bmatrix} h(X_1) \\ h(X_2) \\ \vdots \\ h(X_N) \end{bmatrix} = \begin{bmatrix} g(\omega_1^\tau X_1 b_1) & \cdots & g(\omega_L^T X_1 b_L) \\ \vdots & \ddots & \vdots \\ g(\omega_1^\tau X_N b_1) & \cdots & g(\omega_L^T X_N b_L) \end{bmatrix}_{N\times L} \end{cases} \quad (2)$$

In Formula (2), H represents the hidden layer state. β represents the output layer weight. N means the amount of samples. L expresses the amount of hidden layer nodes. τ denotes the transposition of the objective matrix value. For the input layer weights ω and bias b, their values are usually determined by combining random numbers with activation functions. The minimum loss function is specifically showcased in Formula (3).

$$\begin{cases} \min l = \|Y - H\beta\|_2^2 \\ \hat{\beta} = \vec{H}Y \end{cases} \quad (3)$$

In Formula (3), $\hat{\beta}$ represents the estimation of β training. l represents the loss function. $\vec{H}$ represents the generalized inverse of matrix H. The overfitting risk of RELM increases with the increase of hidden layers. To improve this problem, it is proposed to introduce L_2 regularization term and penalty factor in RELM, and Formula (3) can be rewritten as shown in Formula (4).

$$\min l = C\|Y - H\beta\|_2^2 + \|\beta\|_2^2 \Rightarrow \begin{cases} \min\limits_{\beta} C\|e\|_2^2 + \|\beta\|_2^2 \\ s.t. Y - H\beta = e \end{cases} \quad (4)$$

In Formula (4), C represents the penalty factor. Then, it further optimizes the above equation by introducing Lagrangian multipliers and constructing the Lagrangian equation. By taking zero partial derivatives, Formula (5) can be obtained.

$$\hat{\beta} = H^\tau \left(H^\tau H + \frac{1}{C} \right)^{-1} Y \quad (5)$$

According to Formula (5), the output function of ELM can be obtained, as shown in Formula (6).

$$f(X) = h(X)\beta = h(X)H^\tau \left(H^\tau H + \frac{1}{C} \right)^{-1} Y \quad (6)$$

Then, the kernel function $K(u,v)$ is introduced, replacing $h(X)$ with a kernel function, and the kernel function matrix is defined as $\Omega_{ELM} = HH^T : \Omega_{ELMi,j} = h(X_i)\cdot h(X_j) = K(X_i, X_j)$. Based on the above formula, the output function can be got as indicated in Formula (7).

$$\begin{aligned} f(X) &= h(X)H^\tau \left(H^\tau H + \frac{1}{C} \right)^{-1} Y \\ &= \begin{bmatrix} K(X, X_1) \\ \vdots \\ K(X, X_N) \end{bmatrix}^\tau \left(\Omega_{ELM} + \frac{1}{C} \right)^{-1} Y \end{aligned} \quad (7)$$

CFA is a recommendation algorithm based on user behavior analysis, which can identify similarities between users based on their historical behavior data, and then predict the current user's behavior based on the behavior of similar users [17]. The combination of these two algorithms can better handle the complexity and diversity of passenger flow data, improve the accuracy and reliability of predictions. Collaborative filtering calculation is based on the project. Firstly, it assumes that the user group set, project set, and evaluation set are U, V, and R, respectively. The user is u_i, the project is v_j, and the i user's evaluation of the j project is r_{ij}. The similarity between the two projects is calculated using the cosine similarity calculation method, as shown in Formula (8).

$$w_{ij} = \frac{\left| N(v_i) \cap N(v_j) \right|}{\sqrt{N(v_i) \square N(v_j)}} \quad (8)$$

In Formula (8), $N(v_i)$ means the amount of users interacting with the project v_i. $N(v_j)$ expresses the amount of users interacting with project v_j. $\left| N(v_i) \cap N(v_j) \right|$ represents the number of users interacting with both project v_i

and project v_j simultaneously. Co-occurrence matrix is a commonly used form of data representation that can effectively capture the relationship between users and projects. In the co-occurrence matrix, each row represents a user, each column represents an item, and the elements in the matrix represent the frequency or intensity of interaction between the corresponding user and item. Through this transformation, high-dimensional interaction information can be compressed into a two-dimensional matrix, thereby reducing the complexity of the data [18]. The interaction between users and projects converted into a co-occurrence matrix C, as indicated in Fig. 2, which is a schematic diagram of the co-occurrence matrix transformation.

With the co-occurrence matrix C, the predicted score of user evaluations is calculated, as shown in Formula (9).

$$P_{uj} = \sum_{N(u)\cap S(j,k)} r_{ui} \times w_{ij} \tag{9}$$

In Formula (9), r_{ui} represents the user's true rating of the project, and $S(j,k)$ represents the k projects with extremely high similarity to the project. Then, based on the predicted score, the projects are arranged, and the top ranked projects are recommended to users.

B. Construction of Passenger Flow Prediction Model based on MPA

To solve complex data problems and effectively predict tourist traffic, this study chooses to use the MPA. The MPA is a novel type of biomimetic algorithm that simulates the behavior of marine predators, such as reproduction, migration, and predation, to achieve global search and local fine search of the problem solution space [19]. This algorithm has good global optimization ability and stability, and can effectively handle complex problems such as multi-objective and multi-constraint. Therefore, by using the MPA, the efficiency and accuracy of problem solving can be effectively raised, thereby meeting the accuracy requirements of human flow prediction. In the initial stage of MPA, the amount of search agents is first set to n, the dimension of the variable is set to d, the upper bound of the variable is set to $UB = [ub_1, \cdots, ub_d]$, and the lower bound of the variable is set to $LB = [lb_1, \cdots, lb_d]$. If the prey matrix A consists of all search agents, it will initialize the j dimension of the i th agent, as shown in Formula (10).

$$A_{ij} = r \cdot (ub_j - lb_j) + lb_j \tag{10}$$

In Formula (10), r represents a random number, $r \in [0,1]$. If the fitness function is $fitness(\cdot)$, then the fitness of the corresponding search agent A_i can be expressed as $fitness(A_i)$. The optimal search agent is set as A^*. The optimal agent is repeated and an elite matrix is constructed, with the elite matrix as E and the maximum amount of iterations set as T. Then, for the position update stage of the MPA algorithm, it can be divided into three scenarios based on the different number of iterations t. In the first scenario, the predator's speed is faster than the prey. At this moment, $t < T/3$, the main purpose of this scenario is to explore, which can be expressed as Formula (11).

$$\begin{cases} D_i = R_B \otimes (E_i - R_B \otimes A_i^t) \\ A_i^{t+1} = A_i^t + P \cdot R \otimes D_i, i = 1, \cdots, n \end{cases} \tag{11}$$

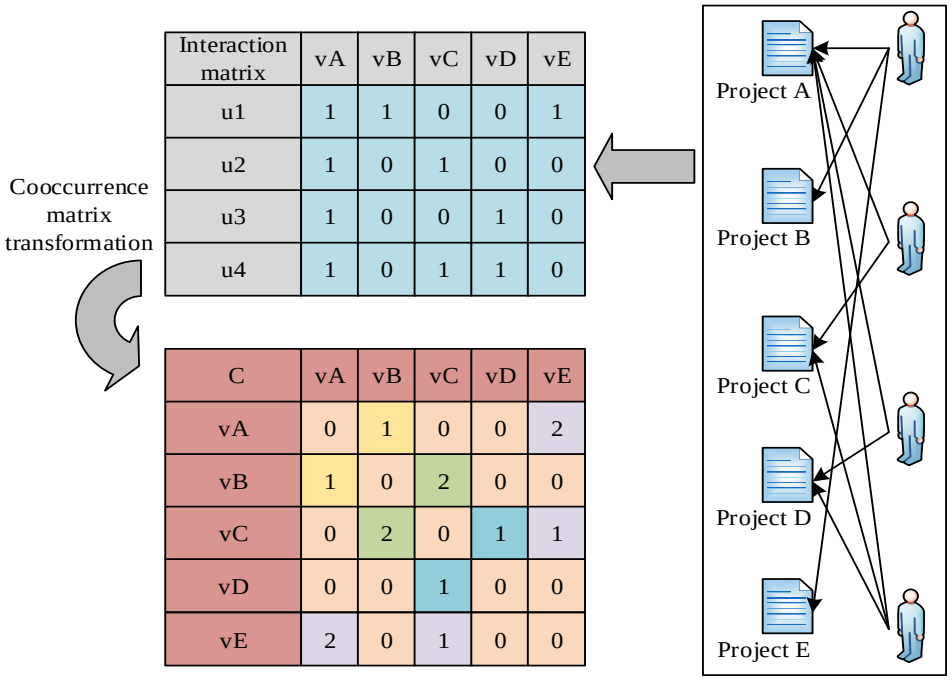

Fig. 2. Schematic diagram of the transformation of the co-occurrence matrix.

In Formula (11), R_B represents the random vector of Brownian motion in the MPA. $\otimes$ represents item by item multiplication. P represents a fixed constant. R represents a uniformly distributed random vector, $R \in [0,1]$, and D_i represents the movement step of the i th predator. In scenario two, the predator and prey move at the same speed. At this point, $T/3 \le t < 2T/3$, the main purpose of this scenario is to explore and simultaneously transition to development. The core idea at this moment is to divide the agent equally for development, as shown in Formula (12).

$$\begin{cases} D_i = R_L \otimes \left(E_i - R_L \otimes A_i^t\right) \\ A_i^{t+1} = A_i^t + P \cdot R \otimes D_i, i = 1, \cdots, n/2 \end{cases} \\ or \begin{cases} D_i = R_B \otimes \left(R_B \otimes E_i - A_i^t\right) \\ A_i^{t+1} = E_i + P \cdot CF \otimes D_i, i = n/2, \cdots, n \end{cases} \tag{12}$$

In Formula (12), R_L represents the Levy motion random number used to simulate the movement of prey, and CF represents the movement amplitude of the predator's motion step D_i. In scenario 3, the speed of the predator is slower than that of the prey. At this moment, $t \ge 2T/3$, the main purpose is to improve the search ability, as shown in Formula (13).

$$\begin{cases} D_i = R_L \otimes \left(R_L \otimes E_i - A_i^t\right) \\ A_i^{t+1} = E_i + P \cdot CF \otimes D_i, i = 1, \cdots, n \end{cases} \tag{13}$$

The flowchart of the MPA is shown in Fig. 3.

In these three scenarios, there will be a problem of slow convergence speed in the initial stage and fast convergence speed in the later stage. At the same time, the lack of more communication between the populations will result in poor diversity performance of the later solutions. Therefore, the study chooses to improve it, reduce the probability of prey random generation, and increase the convergence speed in the early stage. The updated step size after modification is shown in Formula (14).

$$\begin{aligned} D_i &= R_B \cdot \left(Location_D - Location_A\right) \\ or D_i &= R_L \cdot \left(Location_D - Location_A\right) \end{aligned} \tag{14}$$

In Formula (14), $Location_D$ represents the current predator position, and $Location_D$ represents the current prey position. In the study, a method based on boundary crossing is used to construct weights, set reference points, and select individuals to further ensure the diversity and uniform distribution of the population, as shown in Fig. 4.

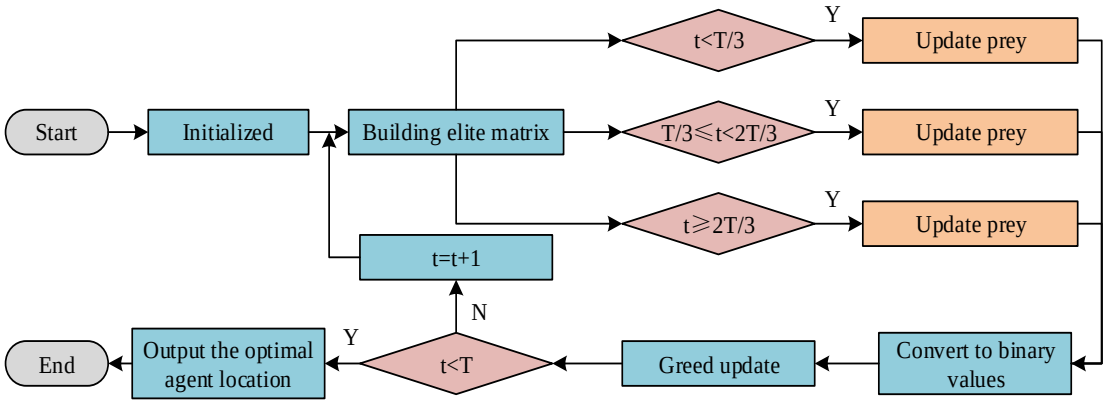

Fig. 3. Schematic diagram of the flow of the marine predator algorithm.

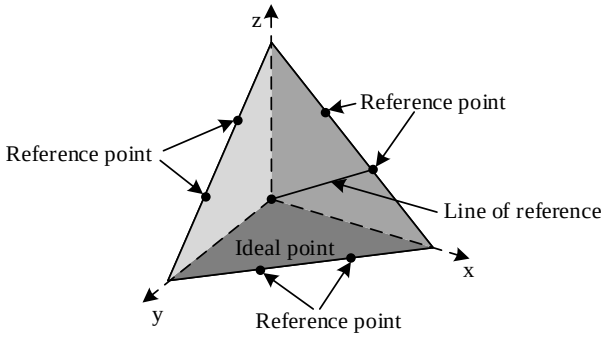

Fig. 4. Schematic diagram of individual selection based on the weight setting reference point of boundary crossing construction.

On the normalized hyperplane in Fig. 4, the target on each dimension is evenly divided into three parts, resulting in 10 reference points. These reference points are evenly distributed on the hyperplane, and this method can make the selected initial population evenly distributed on the real Pareto plane, thereby improving the diversity of the MPA and enhancing its performance.

C. Optimization Design of Multi-strategy Predator Algorithm

To raise the search efficiency and quality of the MSP algorithm, this study chooses to introduce an adaptive adjustment strategy. The adaptive adjustment strategy can dynamically adjust the search strategy based on the complexity of the problem and the search performance of the algorithm, making the algorithm more adaptable to the characteristics of the problem, thereby effectively improving the search efficiency and quality of the solution [20]. In addition, the adaptive adjustment strategy can also improve the robustness of the algorithm, enabling it to perform well in different problems and environments. The study uses two nonlinear degradation

functions, F_1 and F_2, to optimize it, as shown in Formula (15).

$$\begin{cases} F(t)=\dfrac{1}{T} \\ F_1(T)=1-F^{a} \\ F_2(T)=1-F^{1/a} \end{cases} \tag{15}$$

In Formula (15), $a=2/3$ means the iteration times, and T denotes the max amount of iterations. It is further considered based on different scenarios. Take the random number generated during the algorithm optimization process as the object, and when it is less than F_2, it will proceed to Scenario 1, which is the exploration phase. When it is greater than F_2 and less than or equal to F_1, it will proceed to Scenario 2, which is the exploration and transition phase. If it is greater than F_1, it will proceed to scenario 3 and proceed to the development phase. Finally, the process of using the MPA to filter the optimal features in the study is shown in Fig. 5.

As shown in Fig. 5, this is the flowchart of using the MPA to filter the optimal features. In this process, it first initializes and sets the number of agents and iterations. Then, it will generate a subset of features and evaluate them. It is substituted into the model for training and testing the quality of the feature subset. Afterwards, it selectively updates the feature subset according to the situation. Termination condition judgment: If the maximum iteration number is satisfied, it will output the current optimal feature, otherwise, return for reevaluation. Finally, a prediction model is constructed using the optimal output features and the test set is used for prediction. The overall process of the MSP algorithm model for PFP ultimately constructed in the study is shown in Fig. 6.

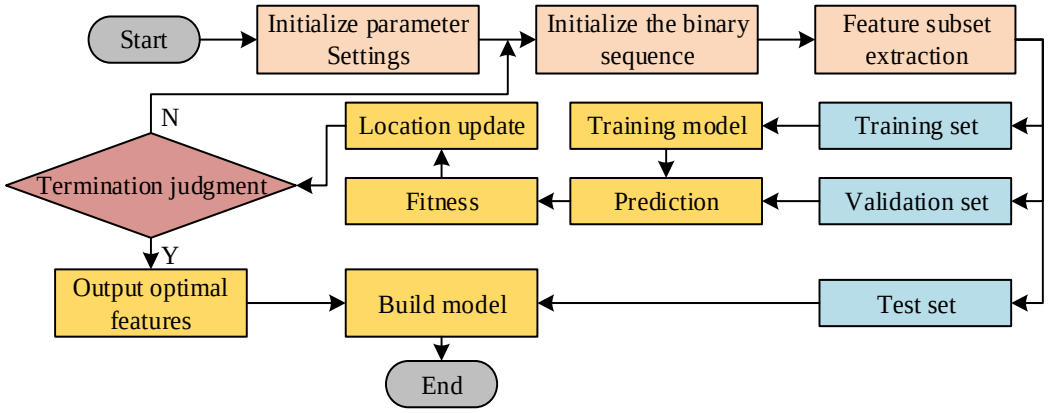

Fig. 5. Flow chart of the best feature selection based on MPA.

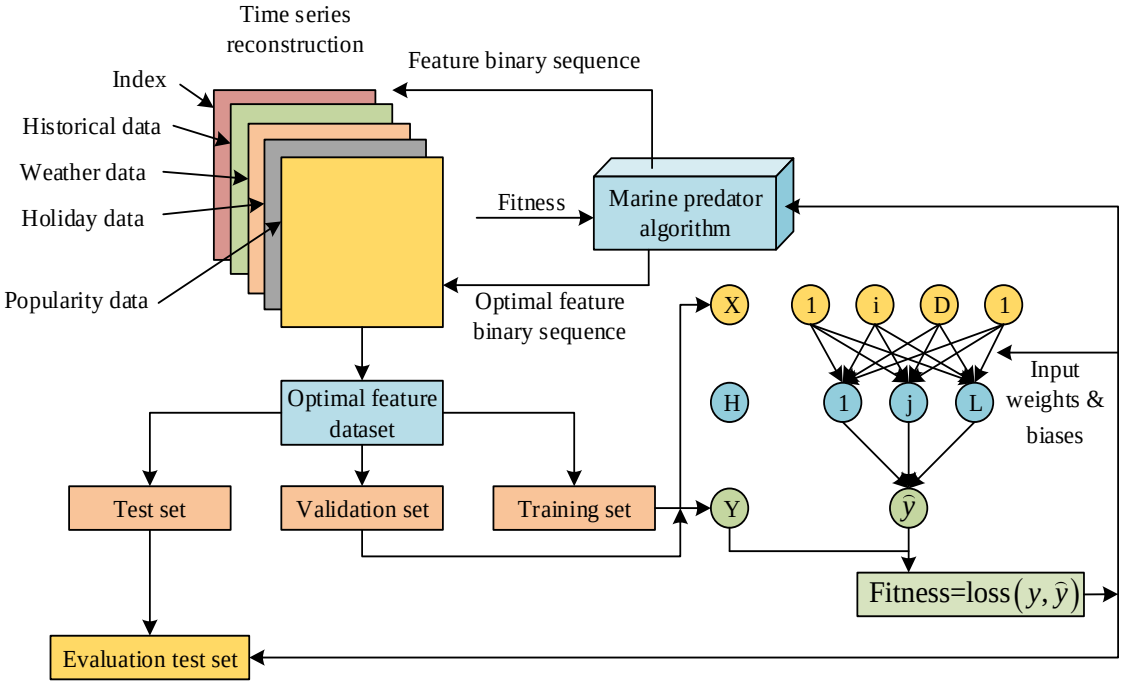

Fig. 6. Flow diagram of multi-strategy predator algorithm for passenger flow prediction.

As shown in Fig. 6, the overall flowchart of a MSP algorithm model for PFP is presented. In this process, time series reconstruction is first combined to process the multi-layer information in the obtained data for subsequent model use. Then a MSP algorithm is applied to filter and obtain important feature information. Finally, the parameters of the MSP algorithm are optimized through RELM. By introducing regularization terms, it is possible to effectively prevent overfitting of the model and ensure its stability. Meanwhile, the training process of RELM is very fast, which can greatly raise the computational effectiveness of the model. The final construction of an accurate and reliable PFP model not only accurately predicts future passenger flow, but also has good stability and computational performance, providing a powerful tool for actual traffic management and planning.

IV. Validation and Testing of Big Data Multi-strategy Predator Algorithms

To test the usability and related performance of the big data MSP algorithm proposed by the research, with the consent of relevant departments, relevant passenger flow data of a certain scenic area from 2022 to 2023 were obtained. The data included characteristics such as total passenger flow, the ratio of domestic and international passenger flow, and the difference between peak and off-season travel flows, and combined information such as the age distribution of passengers, gender ratio, booking channel preferences, and the diversity of travel destinations. Data cleaning was performed, outliers were removed, and normalization was performed to eliminate the effects of different dimensions. 80% of this dataset was utilized for training and the remaining 20% for testing. The study further introduced LSTM, RF, and SVR to compare with the proposed methods. Due to the large size of the dataset, to avoid hardware performance affecting the performance of the model, the study chose to rent a cloud server platform for testing. When conducting research on PFP, it may be necessary to process a large amount of data and run complex machine learning models, thus requiring a powerful cloud server platform. The specific hardware, software, and training parameter settings are indicated in Table I.

To ensure the validity and reliability of the comparison results, the implementation details and parameter settings of each method in the experiment were referred to the original literature, and were transparently and consistently applied in the study. All experiments were repeated under the same hardware and software environment to ensure repeatability of results. To further enhance the reliability of the comparison results, the study invited experts in the field to review the experimental design and results, and only the reliable results were retained. The convergence performance of the four models was tested, with specific test indicators being the relative values of F1 and Recall. The test results are denoted in Fig. 7. From Fig. 7(a), the proposed MSP algorithm reached its optimal state around the 30th iteration, with an F1 value of 0.846, which was 0.124-0.362 ahead of the other three models. In Fig. 7(b), the proposed method had the best convergence speed, and its recall value was 0.862, which was 0.117-0.389 ahead of the other three models.

TABLE I. Hardware and Software Details and Training Parameter Settings

Hardware			Software		Training parameter	
Name	**Supplier**	**Details**	**Name**	**Details**	**Name**	**Details**
Amazon Web Services	Amazon		OS	Amazon Linux 2 AMI	Optimizer	Adam
Instance type	Amazon	c5.9xlarge	Python	3.8.10	Learning rate	0.001
CPU	Intel	Xeon Platinum 8000	MySQL	8.0.23	Batch size	64
vCPU	-	36 core	Apache Hadoop	3.2.1	Epochs	100
RAM	-	32 GB	Apache Spark	3.1.1	Gradient clipping	5
MEM	-	900 MB/s	TensorFlow	2.4.1	Regularization	L2-0.1
Network	-	Elastic network adapter	Keras	2.4.3	Dropout	0.5

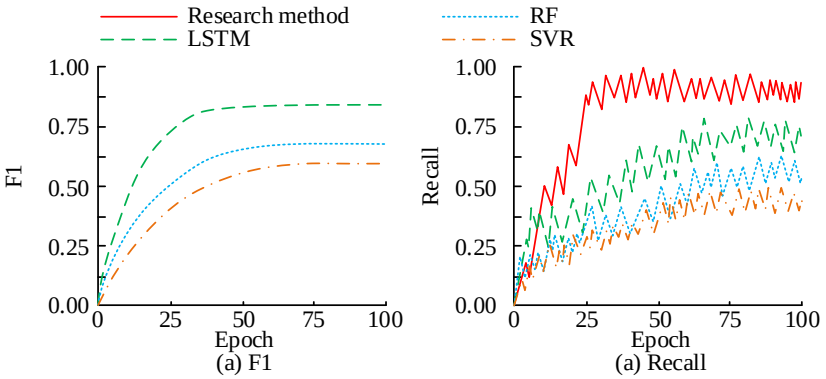

Fig. 7. F1 value and Recall value test results of four models.

The comprehensive performance indicators of four models were tested, including mean absolute error (MAE), root mean square error (RMSE), coefficient of determination (R^2), normalized root mean square error (NRMSE), and average absolute percentage error (MAPE). To minimize the impact of error, each model was tested three times, and the test outcomes are denoted in Table II. From Table II, all performance indicators of the method proposed by the research performed well, with an MAE value of 1858.530, which was the smallest numerical performance, indicating that the proposed method had the best accuracy. The R^2 of the method proposed by the research was 0.9553, which was the highest and closest to 1, indicating that the proposed method had the strongest usability in practice.

The fitting degree of the four models was tested, and the test outcomes are expressed in Fig. 8. From Fig. 8(a), the proposed MSP algorithm had the highest fitting performance, with a fitting degree of 97.8%, which was 6.27%-19.31% higher than the other three models. However, from Fig. 8 (b), 8 (c), and 8 (d), the fitting performance of the other three models was not as good as that of the proposed MSP algorithm.

The PFP results of the four models were tested, as shown in Fig. 9. From Fig. 9, the MSP algorithm proposed by the research could more accurately predict human traffic, with the minimum difference between the predicted and actual values, the highest accuracy, and the lowest error rate. The effectiveness of the research method in capturing complex patterns and associations in the data was demonstrated, thanks to the effectiveness of the RELM in the algorithm, which reduces the risk of overfitting of the model.

TABLE II. MULTIPLE PERFORMANCE TEST RESULTS FOR FOUR MODELS

Model	Time	MAE	MRMSE	R^2	NRMSE	MAPE
Research method	1	1859.627	2645.269	0.9567	19.184	14.186
	2	1854.298	3644.699	0.9639	19.014	15.364
	3	1861.664	3651.894	0.9553	18.624	14.629
LSTM	1	2054.925	2864.193	0.9217	20.641	16.294
	2	2052.815	2864.237	0.9154	21.981	17.262
	3	2051.268	2869.987	0.9036	20.639	16.397
RF	1	2314.148	2968.167	0.9053	23.948	19.636
	2	2315.624	2965.955	0.8751	22.194	18.952
	3	2316.856	2969.330	0.8925	23.962	19.682
SVR	1	2649.854	3012.354	0.8714	24.681	20.018
	2	2657.593	3011.947	0.8659	23.687	21.362
	3	2651.394	3010.492	0.8514	23.591	20.697

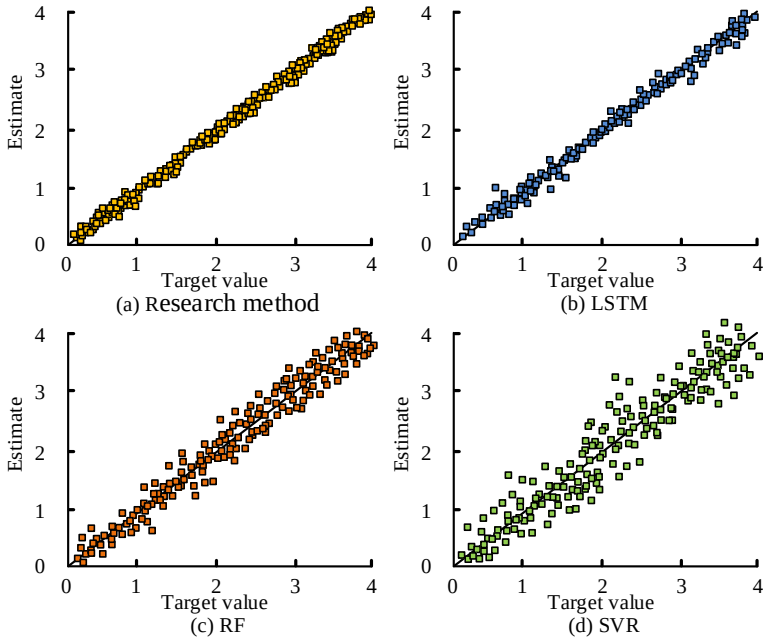

Fig. 8. The fitting test results of four models.

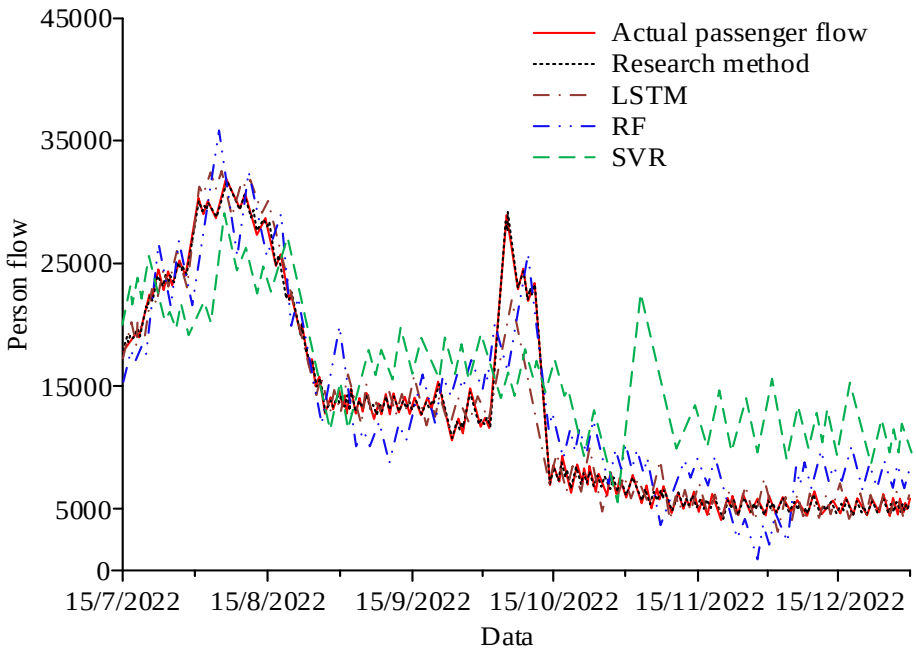

Fig. 9. The prediction results of the person flow of four models.

The actual PFP results of four models were tested. To ensure the objectivity and scientificity of the experiment, a 3-day experiment was conducted, mainly focusing on predicting the actual uplink, downlink, and total flow. The test findings are expressed in Table III. From Table III, the overall error rate of the MSP algorithm proposed by the research was 2.29%, which was 3.47%, 5.53%, and 6.50% higher than LSTM, RF, and SVR, respectively, indicating that it had the highest accuracy in actual PFP and the error met the requirements for practical use. The improvement is due to the efficient search mechanism of the research method, which can quickly guide the algorithm towards the global optimal solution, demonstrating the effectiveness of the research method in capturing complex patterns and correlations in the data. The results show that the method can better fit the actual distribution of passenger flow data and provide a more accurate model for prediction.

TABLE III. THE PREDICTION RESULTS OF THE FOUR MODELS

Days	Models	Actual situation			Prediction situation		
		Up	Down	Total	Up	Down	Total
1	Research method	4512	5341	9853	4505	5098	9603
	LSTM				4715	5026	9741
	RF				4015	4852	8867
	SVR				4915	4968	9883
2	Research method	2516	3456	5972	2612	3452	6064
	LSTM				2745	3615	6360
	RF				2214	3215	5429
	SVR				2014	3155	5169
3	Research method	5378	6123	11501	5462	6021	11583
	LSTM				5145	5902	11047
	RF				5499	6357	11856
	SVR				4982	5816	10798
Rate of deviation (%)	Research method	2.29%					
	LSTM	5.76%					
	RF	7.82%					
	SVR	8.79%					

In summary, the MSP algorithm proposed by the research had the best performance and showed excellent performance in practical use testing. It has high prediction accuracy and lower error rate, and can more accurately predict tourist traffic, providing valuable passenger follow prediction data for local relevant departments, with higher practicality. To further determine the scalability of the research methodology, the study used two different data sets for evaluation. Dataset A contains the monthly passenger flow data of a first-tier city from 2019 to 2021, a total of 36 months, involving the number of passengers, tourism income, holiday distribution and other characteristics. Dataset B is the weekly passenger flow data of the city between 2020 and 2022, a total of 120 weeks of data, including the number of passengers, weather conditions, major events, etc. In Dataset A, the proposed algorithm showed good prediction performance, with MAE value of 1858.530 and R^2 of 0.9553. On Dataset B, the algorithm also performed well, with MAE reduced to 1500.000 and R^2 increased to 0.9650. The results showed that with the increase of data volume and the change of data granularity, the proposed algorithm can maintain high prediction accuracy. To evaluate the scalability of the algorithm, tests were performed on datasets of different sizes and characteristics. Dataset A and Dataset B have a big difference in data volume, which respectively represent the PFP under different time granularity. The experimental results showed that the proposed algorithm can predict effectively on both monthly and weekly datasets, which proves the adaptability and scalability of the algorithm for different data sets.

V. Conclusion

To achieve accurate PFP and provide practical and valuable reference data for planning and judgment of relevant departments, an MSP algorithm was proposed based on the current big data background. The aim was to achieve efficient and accurate PFP through the combination of multiple improved strategies. The experimental results showed that the algorithm could reach its optimal state after 30 iterations, demonstrating excellent convergence performance. At the same time, it performed well in all performance indicators, with an MAE value of 1858.530 and R^2 of 0.9553. Its fitting degree was 97.8%, surpassing the other three models in the comparison of the four models by 6.27%-19.31%. In actual PFP, its error rate was 2.29%, which was 3.47%, 5.53%, and 6.50% higher than LSTM, RF, and SVR, respectively. The above test results fully demonstrated the effectiveness and superiority of the algorithm in PFP. However, the MSP algorithm is very sensitive to parameter settings, and improper parameter selection may affect the accuracy of prediction results. In addition, the performance of this method in dealing with more complex data needs to be considered, and its performance in handling non-linear and high-dimensional data still needs to be improved. In future research, the parameter settings and ability to process nonlinear data of this method can be further optimized, and its adaptability to complex data can be strengthened to improve its predictive performance. It is also looked forward to seeing the wider application of this method in fields other than PFP.

Fundings

The research is supported by: Research topic of the 2023 Henan Social Science Federation: Research on the Integrated Development of Red Flag Canal Spiritual Culture and Tourism from the Perspective of Rural Revitalization (SKL-2023-2025).

References

[1] Zamani E D, Smyth C, Gupta S, Dennehy, D. Artificial intelligence and big data analytics for supply chain resilience: a systematic literature review. Annals of Operations Research, 2023, 327(2): 605-632.

[2] Alkhatib A W, Valeri M. Can intellectual capital promote the competitive advantage? Service innovation and big data analytics capabilities in a moderated mediation model. European Journal of Innovation Management, 2024, 27(1): 263-289.

[3] Islam M A, Jantan A H, Yusoff Y M, Chong C W, Hossain M S. Green Human Resource Management (GHRM) practices and millennial employees' turnover intentions in tourism industry in malaysia: Moderating role of work environment. Global Business Review, 2023, 24(4): 642-662.

[4] Maroufkhani P, Iranmanesh M, Ghobakhloo M. Determinants of big data analytics adoption in small and medium-sized enterprises (SMEs). Industrial Management & Data Systems, 2023, 123(1): 278-301.

[5] Wang J, Yang Y, Wang T, Sherratt R S, ZHang J. Big data service architecture: a survey. Journal of Internet Technology, 2020, 21(2): 393-405.

[6] Sevtsuk A. Estimating pedestrian flows on street networks: revisiting the betweenness index. Journal of the American Planning Association, 2021, 87(4): 512-526.

[7] Cooper C H V, Harvey I, Orford S, Chiaradia A J. Using multiple hybrid spatial design network analysis to predict longitudinal effect of a major city centre redevelopment on pedestrian flows. Transportation, 2021, 48(2): 643-672.

[8] Togashi F, Misaka T, Löhner R, Obayashi S. Application of Ensemble Kalman Filter to Pedestrian Flow. Collective Dynamics, 2020, 5(1): 467-470.

[9] Zhang J, Chen F, Cui Z, Guo Y, Zhu Y. Deep learning architecture for short-term passenger flow forecasting in urban rail transit. IEEE Transactions on Intelligent Transportation Systems, 2020, 22(11): 7004-7014.

[10] Quan R, Zhu L, Wu Y, Yang Y. Holistic LSTM for pedestrian trajectory prediction. IEEE transactions on image processing, 2021, 30(1): 3229-3239.

[11] Ramezani M, Bahmanyar D, Razmjooy N. A new improved model of marine predator algorithm for optimization problems. Arabian Journal for Science and Engineering, 2021, 46(9): 8803-8826.

[12] Ahn I, Yoon C. Global well-posedness and stability analysis of prey-predator model with indirect prey-taxis. Journal of Differential Equations, 2020, 268(8): 4222-4255.

[13] Ghanbari B, Djilali S. Mathematical and numerical analysis of a three-species predator-prey model with herd behavior and time fractional-order derivative. Mathematical Methods in the Applied sciences, 2020, 43(4): 1736-1752.

[14] He X, Zhao X, Feng T, Qiu Z. Dynamical behaviors of a prey-predator model with foraging arena scheme in polluted environments. Mathematica Slovaca, 2021, 71(1):235-250.

[15] Bortuli J A, Maidana N A. A modified Leslie–Gower predator–prey model with alternative food and selective predation of noninfected prey. Mathematical Methods in the Applied Sciences, 2021, 44(5): 3441-3467.

[16] Sun W, Wang Y. Prediction and analysis of CO 2 emissions based on regularized extreme learning machine optimized by adaptive whale optimization algorithm. Polish Journal of Environmental Studies, 2021, 30(3): 2755-2767.

[17] Bag S, Dhamija P, Luthra S, Huisingh D. How big data analytics can help manufacturing companies strengthen supply chain resilience in the context of the COVID-19 pandemic. The International Journal of Logistics Management, 2023, 34(4): 1141-1164.

[18] Kar A K, Kushwaha A K. Facilitators and barriers of artificial intelligence adoption in business–insights from opinions using big data analytics. Information Systems Frontiers, 2023, 25(4): 1351-1374.

[19] Abd Elminaam D S, Nabil A, Ibraheem S A, Houssein E H. An efficient marine predators algorithm for feature selection. IEEE Access, 2021, 9(1): 60136-60153.

[20] Bi Z, Jin Y, Maropoulos P, Zhang W J, Wang L. Internet of things (IoT) and big data analytics (BDA) for digital manufacturing (DM). International Journal of Production Research, 2023, 61(12): 4004-4021.