ISSN: 2472-3703 DOI: 10.6911/WSRJ.202504_11(4).0013

Cargo Volume Prediction Based on K-means Clustering and Back Propagation Neural Network

Zhexi Yu[1, a, *], Zhaoji Zhu[1, b], and Meiping Liao[2, c]

[1]School of Electronic and Information Engineering, Guangdong Ocean University, Zhanjiang, Guangdong, 524088, China

[2]School of Economics, Guangdong Ocean University, Zhanjiang, Guangdong, 524088, China

[a]2936083743@qq.com, [b]2240584740@qq.com, [c]2806980812@qq.com

[*] Corresponding author

Abstract

Volume forecasting for sorting centers in e-commerce logistics networks is important for optimizing resource scheduling and transportation efficiency. In this paper, a joint prediction method incorporating K-means clustering algorithm and Back Propagation neural network (BP neural network) is proposed for 57 sorting centers after route adjustment. Firstly, based on the historical transportation data, we construct the logistics network topology map and identify the key hub nodes, and then divide the sorting centers into 3 classes by K-means algorithm to reveal the characteristics of their cargo volume distribution. Then, combined with the information of future transportation route changes, the training and prediction sets of daily and hourly cargo volume are constructed, and the BP neural network model is used for multi-scale prediction. Finally, Analysis of Variance (ANOVA) further shows that the clustering results are statistically significant, indicating that there are significant differences in the shipment volumes of different categories of sorting centers. In addition, in order to test the accuracy of the BP neural network, this paper conducted mean square error and regression analysis to verify the reliability of the prediction results. The model provides a scientific basis for personnel scheduling and transportation optimization in sorting centers, and has high robustness and practical value.

Keywords

Volume forecasting; K-means clustering algorithm; Back Propagation neural network; analysis of Variance; Regression analysis.

1. INTRODUCTION

1.1. Problem Background

The e-commerce logistics network[1] consists of several links in order fulfillment. Figure 1 is a simple logistics network schematic diagram. Among them, the sorting center as an intermediate link in the network, how to efficiently sort the flow plays a very important role in the fulfillment cost and operation cost of the overall network. Accurate cargo prediction for the sorting center is an important issue in the whole logistics network. Forecasting the volume of goods is the basis of management, if you can accurately predict the flow of goods, you can do a good job in advance of the allocation of resources to achieve cost savings and improve efficiency.

At the same time, the prediction of the volume of goods in the sorting center is related to the network transport lines between the sorting centers, and by analyzing the transport volume of

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202504_11(4).0013

each line, it is possible to derive the network connection relationship between the sorting centers. When successfully predict the sorting center of the amount of goods, the scheduling of personnel become a problem. The employees of the sorting center are divided into regular and temporary workers, and the reasonable arrangement of personnel according to the results of cargo volume prediction can reduce the labor cost as much as possible. Therefore, through scientific and reasonable algorithms and modeling to obtain the predicted amount of goods, for e-commerce logistics network system is of great help, can effectively improve efficiency and reduce costs.

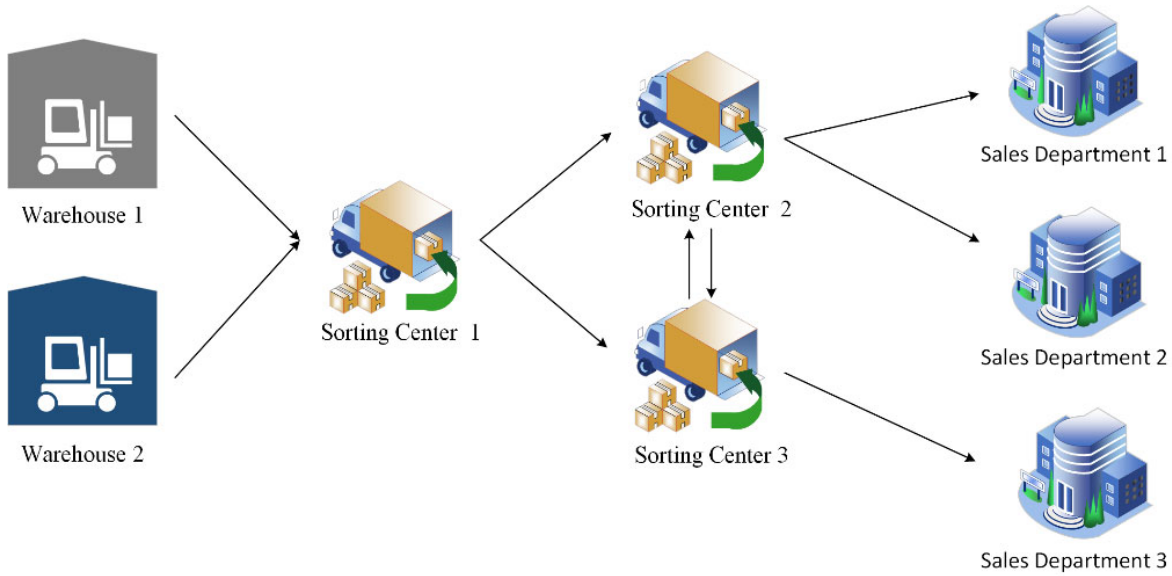

Figure 1. Schematic diagram of logistics network

1.2. Problem Formulation

For all sorting centers, each day is divided into six shifts: 00:00-08:00,05:00-13:00,08:00-16:00,12:00-20:00,14:00-22:00,16:00-24:00, and each personnel can only attend one shift per day.

The logistics network consists of 57 sorting centers, and we know the daily cargo volume of each sorting center in the past 4 months, the hourly cargo volume in the past 30 days, the average cargo volume of each transportation route between sorting centers in the past 90 days, the form of changes in the transportation routes between sorting centers in the next 30 days, and the forecasts of the daily and hourly cargo volume of each sorting center in the next 30 days before the changes in the transportation routes. The above five types of data are labeled as Data 1, Data 2, Data 3, Data 4, and Data 5. Based on the above data, the following problems need to be solved:

(1) Analyze the relationship between sorting centers.

(2) Forecast the daily and hourly cargo volume for the next 30 days for 57 sorting centers and solve for the forecast.

(3) Evaluate the predicted values and analyze the robustness and accuracy of the model.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202504_11(4).0013

2. PROBLEM ANALYSIS AND MODEL CONSTRUCTION

2.1. Analysis of Question

From the problem background analysis, we know that the sorting nodes present a mesh structure among them. According to Data 3, we can get the transportation line relationship diagram before its unchanged transportation line by constructing a tree. Then use K-means clustering algorithm[2] to get the clustering center and find the center sorting point. According to the degree of the sorting center to divide the class, so as to transport optimization. Then Data 1 and 2 are combined with Data 3, 4 and 5 to construct daily and hourly features respectively and get the training and test sets. The Back Propagation neural network[3] (BP neural network) model is constructed based on the data set obtained above, and the predicted values of daily and hourly cargo volume of 57 sorting centers for the next 30 days can be obtained. In order to verify the fit of the model, we select the predicted values obtained by day and demonstrate them through regression plots, error plots and convergence plots.

2.2. Model Assumptions

(1) It is assumed that changes in routes directly affect the increase or decrease in the number of cargo arrivals and can be determined from historical data.

(2) It is assumed that in the process of cargo transportation, the influence of uncertain natural factors such as weather, temperature, and social factors is ignored.

(3) It is assumed that there will be no discrepancies in the volume of cargoes due to misdirection of cargoes during transportation.

2.3. Cargo Prediction Modeling for Sorting Centers

2.3.1 Construction of logistics topology maps

$\boldsymbol{W_i}$ is the net supply of sorting center $\boldsymbol{i}$. Since the net outflow of goods from each sorting center is equal to its supply:

$$\sum_{j=1}^{N} X_{ij} - \sum_{k=1}^{N} X_{ki} = W_i \quad \forall i \in \{1,2,...,N\} \tag{1}$$

where $\boldsymbol{X_{ij}}$ is the flow of goods from sorting center $\boldsymbol{i}$ to j $\left(\boldsymbol{X_{ij} \geq 0}\right)$, $\boldsymbol{X_{ki}}$ is the flow of goods from sorting center k to $\boldsymbol{i}$ $(\boldsymbol{X_{ki} \geq 0})$, and $\boldsymbol{i, j, k \in \{1, 2, \ldots, N\}}$.

All cargo flows $\boldsymbol{X_{ij}}$ must satisfy the non-negative condition:

$$X_{ij} \geq 0 \ \forall i, j \in \{1,2,...,N\} \tag{2}$$

2.3.2 Solution of classification of cargo volume between sorting centers based on K-means clustering algorithm

The steps of the algorithm are as follows:

Step1: Data initialization. At the beginning of the algorithm, K data points need to be randomly selected as the initial clustering centers.

Step2: Distribution of data. Assign each data point to the nearest cluster center. The distance value from each data point in the dataset to each cluster center is calculated separately, and then the value is assigned to the closest cluster center. This step uses Euclidean distance as a distance metric and is calculated as follows:

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202504_11(4).0013

$$dist(x, c_i) = \sqrt{\sum_{j=1}^{d} (x_j - c_{ij})^2} \tag{3}$$

where x is the data point, c_i is the ith clustering center, d is the dimension of the data, and x_j and c_{ij} are the values of x and c_i in the jth dimension respectively.

Step3: Update and recalculate the center of each cluster. The center of the new cluster is the mean of all data points within that cluster, calculated as follows:

$$c_i = \frac{1}{|S_i|} \sum_{x \in S_i} x \tag{4}$$

where S_i is the set of data points for the ith cluster and $|S_i|$ is the number of data points in the set.

Step4: iteration. Repeat Step2, Step3 until the centroid of the cluster no longer changes or the change is less than some threshold.

2.3.3 Solving cargo forecasting based on BP neural network

For this model, we only need to find the appropriate weight ratio ω_j and thresholds θ_j, then the model is able to achieve the success of predicting the actual future shipment with the knowledge of the previous data. Thus, the mean square error E_k is shown below:

$$E_k = \frac{1}{2} \sum_{j=1}^{l} \left(y_f^k - y_j^k \right)^2 \tag{5}$$

Let the input β_j of the jth output neuron be:

$$\beta_j = \sum_{h=1}^{q} \omega_{hj} b_h \tag{6}$$

where the hjth weight value is ω_{hj}.

Based on the gradient descent strategy, ω iterations are performed, followed by k iterations. Using Taylor's formula, the function $E(\omega)$ is expanded at ω_k for another $k+1$ rounds of iterations. Let $\omega = \omega_{k+1}$:

$$E(\omega_{k+1}) = E(\omega_k) + E'(\omega_k)(\omega_{k+1} - \omega_k) \tag{7}$$

To make the model error value smaller, there is:

$$E(\omega_{k+1}) - E(\omega_k) = E'(\omega_k)(\omega_{k+1} - \omega_k) < 0 \tag{8}$$

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202504_11(4).0013

The above equation can be made to hold when the special solution $\omega_{k+1} - \omega_k = -E'(\omega_k)$, which is substituted into the above equation:

$$E(\omega_{k+1}) - E(\omega_k) = -2[E'(\omega_k)] < 0 \tag{9}$$

When the values of ω_{k+1} and ω_k are very similar, $R(\omega)$ can be ignored. we therefore introduce a learning rate η , to reduce the degree of deviation:

$$\omega_{k+1} - \omega_k = -\eta E'(\omega_k) \tag{10}$$

Applying the iterative formula and the chain rule based on the above equation, we can get:

$$\frac{\delta k}{\delta w_{hj}} = \frac{\delta E_k}{\delta y_j^k} \cdot \frac{\delta y_j^k}{\delta \beta_j} \cdot \frac{\delta \beta_j}{\delta w_{hj}} \tag{11}$$

Denote g_j as:

$$g_j = -\frac{\delta E_k}{\delta y_j^k} \cdot \frac{\delta y_j^k}{\delta \beta_j} \tag{12}$$

Denote b_h as:

$$b_h = \frac{\delta \beta_j}{\delta \omega_{hj}} \tag{13}$$

Connecting the above equations (10), (12) and (13), we get:

$$\Delta \omega_{hj} = \eta g_j b_h \tag{14}$$

$$\Delta \theta_j = -\eta g_j \tag{15}$$

3. PROBLEM SOLVING AND ANALYSIS OF MODEL ROBUSTNESS

3.1. Analysis of the relationship between unaltered transport routes

The distance of a node from the center and the degree size of the node reflects the degree of connectivity and importance of the node in the network. The higher the degree of a node, the more connected it is to other nodes, with more opportunities for information transfer and communication.

As shown in Figure 2, the node with the largest degree is the sorting node 7 with a degree of 21, which means that node 7 has more connections with other nodes with more opportunities for information transfer and communication, and it is at the core point of the new logistics network. Besides that, the degree of sorting centers 5,8,37 is close to that of sorting node 7, so

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202504_11(4).0013

the freight volume of these sorting centers may be on the high side, and may need a lot of personnel to maintain the turnover of freight volume.

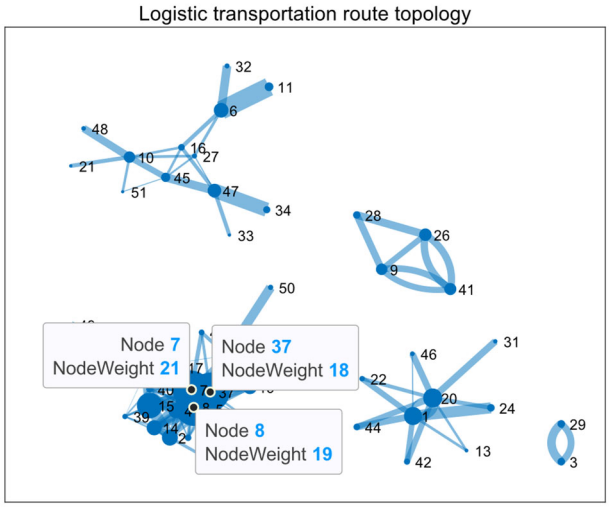

Figure 2. Logistic transportation route topology

3.2. Solving cluster centers based on K-means clustering algorithm

All the nodes involved in cargo transportation are counted by reading Data 3 and cargo flow is calculated for each node. For each node, the total amount of inflow and outflow data is calculated and the larger of the inflow and outflow is selected as the data flow of the node. Eventually, the data set of cargo flow corresponding to each sorting center is derived through integration, as shown in Table 1(due to space limitations, only part of the data is shown in this paper, see Appendix for specific data).

Based on the data in Table 1, clustering calculations were performed on the sorting centers to find the clustering centers shown in Table 2.

Based on the data in Table 2, it is derived to which clustering center each sorting center belongs. The number of sorting centers included in each clustering center is shown in Table 3.

According to Table3, Analysis of Variance[4] (ANOVA) was performed and yielded a significance value of 0. The modeling results are statistically scientific and credible when the significance of ANOVA is < 0.05. Thus, indicating that there is a significant difference in the volume of goods in different cargo centers. And the observed differences between the groups are not due to random factors but reflect the real differences in the overall.

Table 1. Volume data for each sorting center

Sorting center	Volume of cargo
1	**1049**
2	**890**
3	**260**
4	**622**
5	**1321**
6	**1032**
7	**2022**
8	**2174**
9	**413**
10	**592**

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202504_11(4).0013

Table 2. Cargo volume in 3 types of clustering centers

Center of clustering	Volume of cargo
1	**1099**
2	**2386**
3	**345**

Table 3. Number of sorting centers included in each cluster center

category	Amount
1	**11**
2	**3**
3	**43**

3.3. Prediction of future cargo volume based on BP neural network

3.3.1 Prediction of daily cargo volume for the next 30 days

Data 1 is combined with Data 3, 4 and 5 to construct the daily cargo features, thus obtaining the training set and test set. The final obtained daily cargo forecast data for the next 30 days is shown in Table 4 (due to space limitations, only part of the data is shown in this paper)：

3.3.2 Prediction of daily hourly cargo for the next 30 days

Data 2 is combined with Data 3, 4 and 5 to construct the hourly cargo features, and then obtain the training set and test set. The final obtained daily cargo volume forecast for the next 30 days is shown in Table 5 (due to space limitations, only part of the data is shown in this paper).

Table 4. Daily cargo forecast data for the next 30 days

Sorting center	Date	Volume of cargo
1	**2023/12/1**	**51153**
1	**2023/12/2**	**51442**
1	**2023/12/3**	**51733**
1	**2023/12/4**	**52027**
1	**2023/12/5**	**52322**
1	**2023/12/6**	**52619**
1	**2023/12/7**	**52918**
1	**2023/12/8**	**53219**
1	**2023/12/9**	**53521**
1	**2023/12/10**	**53826**

4. MODEL IMPROVEMENT AND EXTENSION

4.1. Accuracy analysis of BP neural networks

When the constructed BP neural network model is trained on the daily cargo volume, the convergence curve of the model and the relevant regression plots of the training data are shown in Figure 3, Figure 4.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202504_11(4).0013

Table 5. Daily hourly cargo forecast data for the next 30 days

Sorting center	Date	Hour	Volume of cargo
1	**2023/12/1**	**0**	**2918**
1	**2023/12/1**	**1**	**2929**
1	**2023/12/1**	**2**	**2924**
1	**2023/12/1**	**3**	**2908**
1	**2023/12/1**	**4**	**2787**
1	**2023/12/1**	**5**	**1786**
1	**2023/12/1**	**6**	**1241**
1	**2023/12/1**	**7**	**1209**
1	**2023/12/1**	**8**	**1233**
1	**2023/12/1**	**9**	**1313**

As shown in Figure 3, the mean square error[5] of each round is almost maintained at 10^7 to 10^8, which indicates the accuracy of the BP neural network prediction model used in this paper. As shown in Figure 4, the correlation coefficients R[6] of the three plots are all greater than 0.9, so the predicted values and the results have a completely positive correlation, indicating that the predicted values are highly related to the actual values. Therefore, the prediction results obtained by this model are highly accurate.

When the constructed BP neural network model was trained on the daily hourly cargo volume, the convergence curve of the model and the correlation regression plots of the training data are shown in Figure 5, Figure 6.

From Figure 5 the mean squared error for each round is almost maintained between 106, indicating the stability of the BP model for hourly prediction. Figure 6 shows that the correlation coefficient R of all the three plots is close to 0.8, indicating that the predicted values are sufficiently related to the actual values, and therefore the results of the hourly BP model prediction are credible.

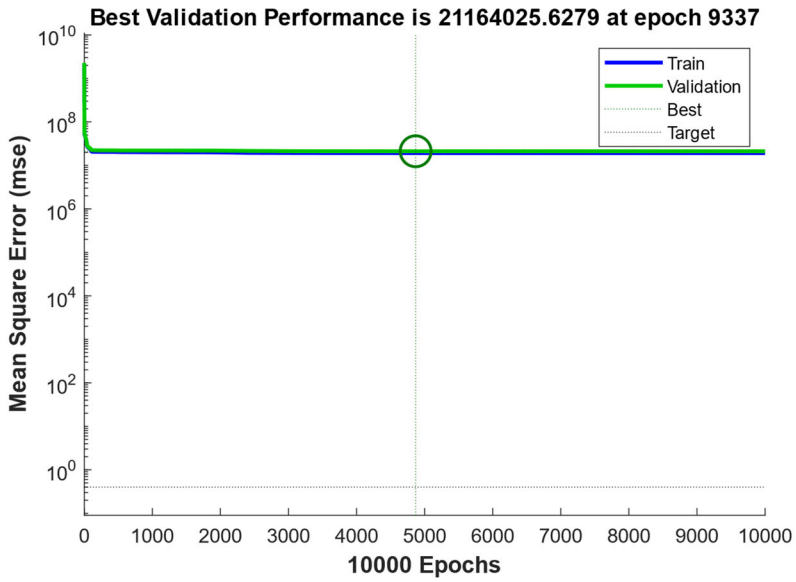

Figure 3. Convergence plot of daily cargo forecast

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202504_11(4).0013

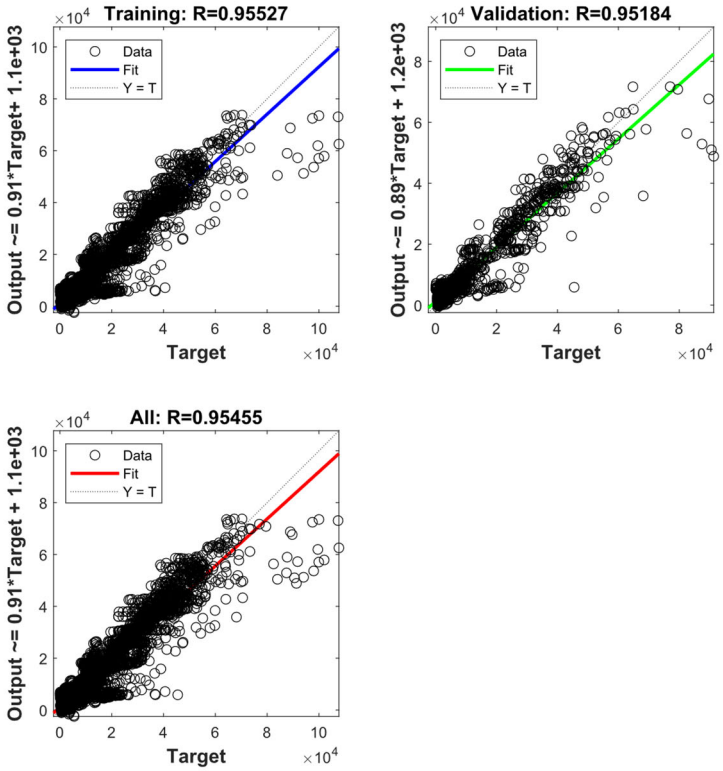

Figure 4. Daily cargo volume forecast regression chart

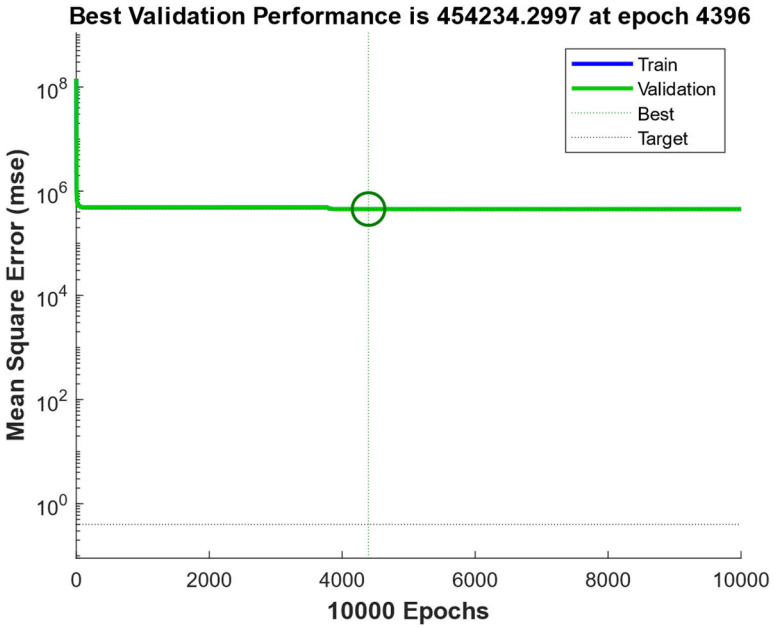

Figure 5. Convergence curves for daily hourly cargo volume data forecasts

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202504_11(4).0013

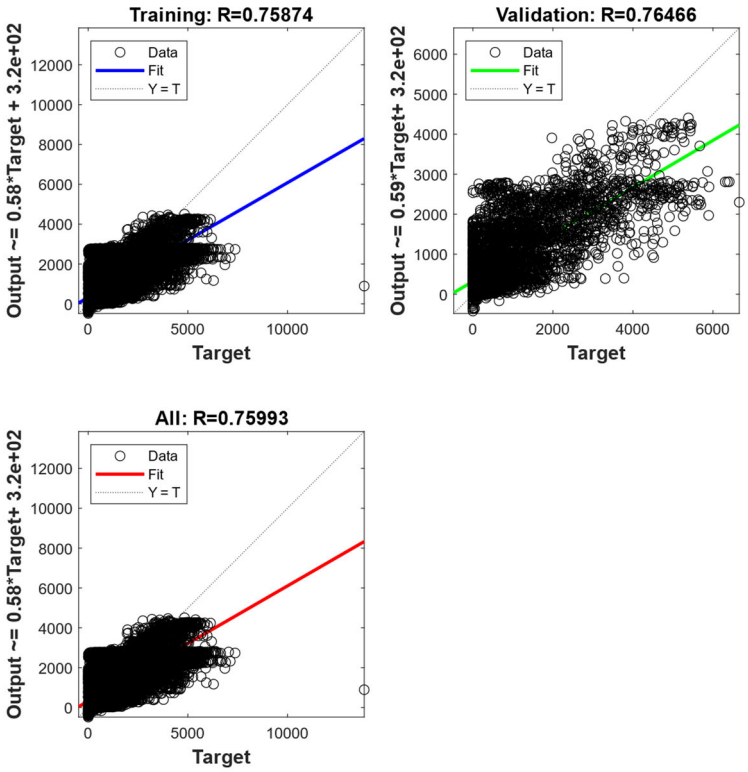

Figure 6. Daily hourly cargo volume forecast regression chart

4.2. Disadvantages of the model

(1) The training phase of BP neural networks is characterized by high data dependency and substantial computational overhead, rendering the model susceptible to overfitting due to insufficient regularization mechanisms.

(2) The K-means clustering algorithm necessitates an a priori specification of the cluster count, which lacks rigorous validation through objective criteria such as silhouette analysis or gap statistics. Furthermore, its performance exhibits sensitivity to initial centroid selection, often resulting in convergence to local optima due to the non-convex nature of the optimization landscape.

4.3. Further improvement and extension of the model

(1) Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) provides a non-parametric clustering framework that supersedes k-means by leveraging local density gradient analysis and mutual reachability metrics to intrinsically infer cluster cardinality, thereby circumventing the heuristic bias associated with manual k-value initialization in centroid-dependent algorithms.

(2) A Spatio-Temporal Cross-Attention Network (ST-CAN) is proposed to explicitly model long-range dependency in freight volume dynamics across distributed nodes through multi-head self-attention mechanisms, thereby addressing the inherent locality bias of BP neural networks characterized by restricted receptive fields and inadequate contextual feature aggregation.

The above improvements will promote the evolution of this model from a single prediction model to an intelligent logistics decision-making system, such as the optimization of freight flow and path planning for multi-hub logistics networks, and the collaborative optimization of

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202504_11(4).0013

intelligent distribution paths in high-density cities, etc., which can ultimately be applied to the global optimization and resilience enhancement of the logistics network in the context of Industry 4.0.

REFERENCES

[1] Sven Winkelhaus and Eric H. Grosse: Logistics 4.0:a systematic review towards a new logistics system,Vol. 58 (2020) No.1, p.18-43.

[2] S. Wang, *et al*: K-Means Clustering With Incomplete Data, IEEE Access, Vol. 7 (2019), p.69162-69171.

[3] Q. Luo: Logistics Demand Forecasting Based on BP Neural Network Model Optimized by Genetic Algorithm, *2023 4th International Conference on Computer, Big Data and Artificial Intelligence* (Guiyang, China, December 15-17), p. 445-450.

[4] Bruno S. Sergi, Vittorio D'Aleo, Sylwia Konecka, Katarzyna Szopik-Depczyńska, Izabela Dembińska, Giuseppe Ioppolo: Competitiveness and the Logistics Performance Index: The ANOVA method application for Africa, Asia, and the EU regions, Sustainable Cities and Society, Vol. 69 (2021), No.102845

[5] S. Amini and S. Ghaemmaghami: Towards Improving Robustness of Deep Neural Networks to Adversarial Perturbations, IEEE Transactions on Multimedia, Vol. 22 (2020) no. 7, p. 1889-1903.

[6] N. Carlini and D. Wagner: Towards Evaluating the Robustness of Neural Networks, *2017 IEEE Symposium on Security and Privacy* (San Jose, CA, USA, May 22-26), p. 39-57