Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 6
ISET 2024

Cargo Volume Forecasting for Logistics Sorting Centers Based on Random Forest Model

Wenfei Zhao [1, *], Xinyang Li [1] and Zhaoxin Li [2]

[1] School of Chemical Engineering, Qingdao University of Science and Technology, Qingdao, China

[1] School of Chemical Engineering, Qingdao University of Science and Technology, Qingdao, China

[2] School of Chemical Engineering, Qingdao University of Science and Technology, Qingdao, China

* Corresponding Author Email: z1582012226@163.com

Abstract. With the rapid development of logistics industry, logistics sorting center, as a key node in e-commerce logistics network, its cargo volume prediction and personnel scheduling strategy are of great significance to improve operational efficiency and reduce costs. The purpose of this study is to deeply analyze the cargo volume data of 57 sorting centers through machine learning method, establish an accurate cargo volume prediction model, and optimize the personnel scheduling strategy based on the prediction results. In this study, we preprocess the historical cargo data, including missing value filling, time series conversion and feature construction. Then, the random forest model was used to carry out fitting analysis on the cargo volume data of different sorting centers. By comparing the evaluation indexes such as mean square error (MSE), root mean square error (RMSE) and mean absolute error (MAE), the superiority of random forest model compared with neural network, support vector regression and linear regression models was verified. This study not only improves the accuracy of cargo volume forecast, but also realizes the reasonable allocation of human resources, and provides scientific decision support for the operation and management of logistics sorting center.

Keywords: Logistics sorting center; Cargo volume forecast; Random forest model; Machine learning.

1. Introduction

With the vigorous development of e-commerce, the logistics industry, as an important support, is experiencing unprecedented changes. As a key node in the e-commerce logistics network, logistics sorting center undertakes the important task of efficiently and accurately sorting and distributing goods from the upstream of the supply chain to the downstream. In this process, the accuracy of cargo volume forecast plays a crucial role in optimizing resource allocation, reducing operating costs, improving customer satisfaction and enhancing market competitiveness [1].

However, the traditional cargo volume forecasting methods rely on empirical judgment and simple statistical analysis, which is difficult to cope with the increasingly complex market demand and changing operating environment. In order to solve this problem, machine learning technology, especially random forest model, is introduced in this study to build a more accurate and efficient cargo volume prediction model [2].

The purpose of this study is to establish a prediction system which can accurately predict the cargo volume of logistics sorting center by analyzing the historical cargo volume data of 57 sorting centers. In this study, we first preprocessed the historical volume data, including missing value filling, time series conversion and feature construction, to ensure the quality and integrity of the data. Then, we used the random forest model to perform fitting analysis on the cargo volume data of different sorting centers, and compared it with several other machine learning models to verify the superiority of the random forest model.

Through the optimization and comparison of the model, this study not only improves the accuracy of the cargo volume forecast, but also realizes the reasonable allocation of human resources, and provides scientific decision support for the operation and management of logistics sorting center. In

addition, this study also discusses the application potential of random forest model in cargo volume prediction of logistics sorting center, which provides a new idea and method for future research in related fields.

2. Model Preparation

2.1. Dataset

In this study, we obtained four detailed data sets, namely: data set containing daily volume of 57 sorting centers in the past four months 1; Data set covering the hourly volume of these sorting centers over the past month 2; Data set recording the average volume of cargo on the transport route between the sorting centers over the past 90 days 3; And a data set containing changes in the routing of the sorting center over the next 30 days 4. These data sets together constitute a comprehensive cargo volume database of logistics sorting center, which provides solid data support for us to further establish cargo volume forecasting model. Through systematic pre-processing and analysis of these data, we are able to gain a deep understanding of the operation mode of the sorting center to achieve accurate volume prediction.

2.2. Data Preprocessing

In the data preprocessing phase, we first conducted a thorough review of daily volume data from 57 sorting centers over the past four months, aiming to identify and fill in missing values and ensure continuity in time series analysis. Dataset 1 and dataset 2 have basically the same volume distribution, for example, the peak volume sorting centers in both data sets are nearly the same. This shows that the model we applied to data set 1 is also applicable to dataset 2. Although the two datasets have nearly the same proportion of total volumes for different sorting centers, there are differences in total volumes. Apart from the fact that dataset 2 has only one month's data, we speculate that the difference may also be caused by certain missing values in dataset 2. Therefore, we carried out counting and visualization processing on the hourly cargo volume of each sorting center in dataset 2, and the results were shown in Fig. 1. Where the horizontal axis represents the time, that is, the different hours of the day, while the vertical axis represents the quantity of goods recorded in each hour.

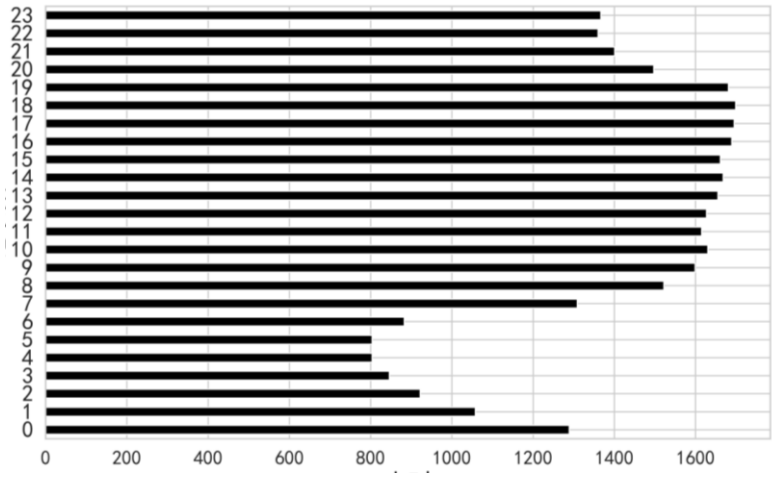

Fig. 1 Volume counts at different hours in Dataset

As can be seen from Fig. 1, the number of sorting varies in each different time interval, especially from 0 to 8, with a large number of missing records, which means that the loss of data for some reason has led to a large deviation in data analysis. In this case, in order not to change the original data distribution law, we need to perform linear interpolation on the missing values. Using linear interpolation, we estimate the missing value in the hourly volume data in dataset 2. All date fields are then uniformly converted into date-time format for subsequent time series analysis.

In addition, in order to enhance the ability of the model to capture the changing trend of cargo volume, we construct a number of features including the trend, lag and moving average of cargo volume before predicting cargo volume. This pre-processing step lays a solid foundation for subsequent machine learning model building and predictive analysis.

3. Construction of Cargo Forecasting Model

In the modern logistics system, the traditional time series prediction model can only produce efficiency in the short term, but it will gradually weaken with the increase of time interval. Especially in the task of long-term prediction, the traditional time series prediction model seems to be powerless. In this paper, it is also unable to take advantage of other relevant factors that can affect the volume of goods. Therefore, we propose to use unconventional time series prediction model to forecast.

3.1. Random Forest Model

Random forest models have a profound background, Boosting method [3] and Boosting method are the first ones to apply ensemble learning algorithm to data classification field of data mining Bagging method [4], aiming at the practical problem that the performance of a single classifier cannot be improved, and referring to the central content of ensemble learning algorithm, puts forward the idea of generating multiple classifiers to reduce the generalization error of the algorithm and thus improve the performance of the algorithm, and applies this idea to the field of data classification. With the development of ensemble learning methods,Tin Kam Ho proposed the concept of Random Decision Forests in 1995 [5], and three years later, he proposed the ensemble method of random subspaces. Breiman systematically elaborated the random forest method, and the random forest algorithm officially became an important part of the data mining classification algorithm. In the past 10 years, due to its excellent performance, random forest algorithm has been widely used in practice, and has become a hot research topic for researchers and technicians in many fields such as data analysis and mining, knowledge management, and pattern recognition.

Random forest (RF) is a combinatorial classifier, which uses bootstrap resampling method to extract multiple samples from the original sample, build a decision tree model for each bootstrap sample, and then combine these decision trees together to get the final classification or prediction result by voting. A large number of theoretical and empirical studies have proved that random forest algorithm has high prediction accuracy, good tolerance to outliers and noise, and is not prone to overfitting. This chapter intends to make a systematic summary and arrangement of the random forest research in order to facilitate the subsequent optimization research.

3.2. Model Optimization

In the aspect of optimization of random forest itself, we conducted a detailed analysis on the factors affecting the classification performance of random forest algorithm. Aiming at the phenomenon of different performance of random forest caused by different node splitting algorithms in the generation process of random forest, we proposed a random forest optimization algorithm based on linear transformation, and this algorithm also has strong performance in machine learning. After the data preprocessing algorithm is integrated into the new algorithm, the corresponding program implementation is carried out, and the classification performance optimization of random forest algorithm is realized well.

3.3. Model Comparison

In this paper, in addition to using the random forest model for cargo prediction, we also use several other machine learning models, including Neural Network, support vector regression (SVR), Support Vector Regression and Linear Regression. The following is a brief description of these models.

3.3.1. Neural network

Neural network is a mathematical model designed to imitate the working mechanism of human brain. It simulates the function of human brain to recognize and process information through a network structure composed of a large number of nonlinear neuron nodes. In the neural network model, the data enters through the input layer, is processed by several hidden layers, and finally gets the prediction result in the output layer. Neural networks learn relationships between data by adjusting weights and biases in the network. Neural networks perform well when dealing with complex and non-linear problems, but can require large amounts of data and computational resources.

3.3.2. Support Vector Regression (SVR)

Support vector regression is a regression version of a support vector machine (SVM) that predicts continuous target values by finding the optimal hyperplane. SVR tries to find a function that fits the relationship between inputs and outputs in the training data set as best as possible within a defined margin of error. SVR is particularly good at handling high-dimensional data and has good generalization ability for nonlinear problems. A key feature of SVR is its regularization term, which helps prevent the model from overfitting.

3.3.3. Linear Regression

Linear regression is a statistical method used to determine the best linear description of the relationship between two or more variables. In machine learning, linear regression is often used to predict the output of a continuous numerical type. A linear regression model attempts to find a linear equation that predicts the dependent variable based on one or more independent variables. Linear regression models are simple, easy to understand and implement, but may not be effective enough when dealing with complex nonlinear relationships.

3.4. Evaluation Metrics

In this paper, we compare the performance of these models in the task of cargo volume forecasting, including mean square error (MSE), root mean square error (RMSE) and mean absolute error (MAE), etc., to evaluate their forecasting accuracy.

In machine learning and statistics, MSE, RMSE, and MAE are commonly used to measure the accuracy of model predictions. The mean square error is defined as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2. \tag{1}$$

Where n is the number of samples, Y_i is the i th observed value (true value), and $\hat{Y}$ is the i predicted value.

RMSE is defined as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}. \tag{2}$$

RMSE is the square root of MSE, and it has the same units as the original data, so you can more intuitively assess the size of the prediction error.

MAE is defined as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right|. \tag{3}$$

MAE calculates the average of the absolute value of the difference between each predicted value and the true value.

4. Results

In the experiment, we evaluate the accuracy and bias of different model predictions by analyzing the residual graph to ensure the accuracy in the model prediction process. We will build and train the predictive model individually for the pass sorting center. In order to achieve a model that accurately captures the unique volume dynamics of each sorting center, the features are then functionally constructed and evaluated using MAE, RMSE and MSE. Some of the results are shown in Table 1.

Table 1. Volume forecast results for some sorting centers

Sorting Center	Index	Randomforest	NeuralNetwork	SVR	LinearRegression
SC5	MAE	0.069498148	0.089095416	0.09520676	0.078185259
	MSE	0.025309972	0.035432409	0.025852929	0.02629238
	RMSE	0.15909108	0.188234984	0.160788459	0.162149253
SC7	MAE	0.083189776	0.132317903	0.096949275	0.092709286
	MSE	0.033304457	0.039212279	0.031443163	0.040380869
	RMSE	0.182495087	0.198020907	0.177322201	0.200949917
SC57	MAE	0.08989349	0.108659211	0.102269361	0.098766639
	MSE	0.03508154	0.035436767	0.026705428	0.028744991
	RMSE	0.187300666	0.18824656	0.163417956	0.169543477
SC34	MAE	0.076123778	0.177152496	0.103394664	0.09378223
	MSE	0.027621218	0.056624639	0.038041671	0.039116085
	RMSE	0.166196323	0.237959323	0.195042742	0.197777869

From Table 1, we can see that random forest is significantly better than the other three algorithms. Then, we calculate the average performance results of MAE, MSE, and RMSE for the algorithm. As shown in Table 2.

Table 2. The average performance results

	RandomForest	NeuralNetwork	SVR	LinearRegression
MAE	0.082453	0.116585	0.105269	0.091165
MSE	0.024568	0.036548	0.025864	0.025946
RMSE	0.015658	0.018569	0.166874	0.168452

Based on the comparison of the predicted results above with the real results, we carry out visualization processing in the form of residual graphs. As shown in Fig. 2.

In each subgraph, the X-axis represents the number of samples, i.e. the order of observations in the data set; The Y-axis represents the residual, which is the difference between the value predicted by the model and the actual value. The purpose of the residuals plot is to assess the predictive accuracy of the model, and ideally the residuals should be distributed randomly with no obvious pattern, meaning that the model has no systematic bias. In practice, for each model, the residuals between the predicted and true values are calculated and these residuals are plotted with the number of samples. If the points in the residual plot are randomly scattered on both positive and negative sides of the Y-axis, and there is no obvious distribution trend, this indicates that the model's predictions are not significantly biased. Conversely, if the residuals show a pattern or trend, such as being systematically above or below zero, this may indicate that the model is failing in some way to capture the true nature of the data.

The results of the analysis in Fig. 2 show that the random forest model performs the best of all the models because its residual distribution looks the most random and does not show a clear pattern, which means that it has high accuracy and reliability in predicting the volume of goods in the sorting center.

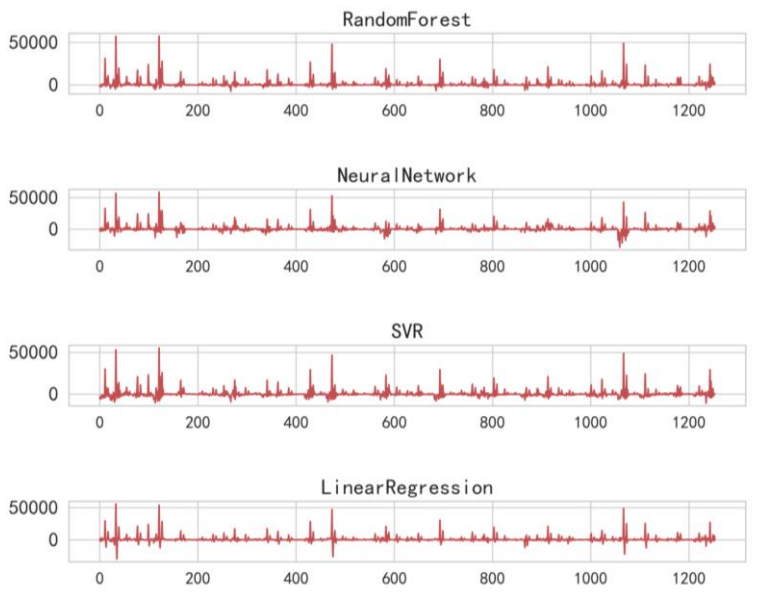

Fig. 2 Residuals of different models

5. Conclusion

In this paper, the machine learning method, especially the random forest model, is used to analyze and study the cargo volume prediction of logistics sorting center. Through careful pre-processing and feature construction of historical cargo volume data from 57 sorting centers, we built an efficient cargo volume prediction model. The results show that random forest model is superior to neural network, support vector regression and linear regression models in forecasting accuracy. In addition, the cargo volume prediction model not only improves the accuracy of prediction, but also provides scientific decision support for the reasonable allocation of human resources and the operation and management of logistics sorting center. By comparing the residual graphs of different models, we further confirm the stability and reliability of the random forest model in the prediction process, and the randomness of the residual distribution shows that the model has no systematic bias.

Despite the positive results of this study, there are some limitations. Future research can be expanded in the following directions: First, more machine learning algorithms and their integration methods c an be explored to find better predictive models; Second, we can consider introducing advanced methods in time series analysis to combine with traditional machine learning models to improve the ability to predict long-term trends. Finally, the applicability and adjustment strategies of the model in different types of logistics sorting centers and different business scenarios can be studied.

References

[1] DianaPerdiguero-Alonso;;FranciscoE.Montero;;AnetaKostadinova;;Juan Antonio Raga;;John Barrett.Random forests, a novel approach for discrimination of fish populations using parasites as biological tags[J].International Journal for Parasitology,2008 (12).

[2] David F. Parkhurst;;Kristen P. Brenner;;Alfred P. Dufour;;Larry J. Wymer.Indicator bacteria at five swimming beaches—analysis using random forests[J].Water Research,2005 (7).

[3] Shi J. Research on optimization of cross-border e-commerce logistics distribution network in the context of artificial intelligence[J]. Mobile Information Systems, 2022, 2022.

[4] Wang Y, Jia F, Schoenherr T, et al. Cross-border e-commerce firms as supply chain integrators: The management of three flows[J]. Industrial Marketing Management, 2020, 89: 72-88.

[5] Leo Breiman.Bagging Predictors[J].Machine Learning,1996 (2).