A Forecast Model based on Two-step Clustering and Random Forest

Xiaojing Wang[1, a], Qingxia Xie[2, b, *], Chuantao Wang[3, c]

[1]School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China

[2]Transportation Informatization Department, China Transport Telecommunications & Information Center, Beijing 100011, China

[3]School of Mechanical-Electronic and Vehicle Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044, China

[a]17120889@bjtu.edu.cn, [b]xieqingxia@cttic.cn, [c]wangchuantao@bucea.edu.cn

*Corresponding author

Keywords: Demand forecast, Two-Step Clustering, Random Forest, and E-commerce

Abstract: Demand forecasting has played an important role in inventory management of e-commerce enterprises in the era of big data. In this study, in order to improve the accuracy of forecasting, a combination model based on the Two-step Clustering and Random Forest is proposed. The Two-step Clustering Algorithm is firstly applied to clustering data series into several disjoint clusters. Then, each cluster is set as the input and output sets to construct the corresponding C-RF model. Finally, the testing set is partitioned into the corresponding cluster by the trained Two-step Clustering model, and then the prediction results are calculated based on the corresponding trained C-RF model. By comparison with the single Random Forest model, the C-RF model based on Two-step Clustering is proved to outperform the single Random Forest model.

1. Introduction

The transition of China's economic structure and the development of the Internet have brought new opportunities for the development of e-commerce. The amount of data from e-commerce enterprises are explosive growth generated in applications and websites of e-commerce [1]. However, big data brings new challenges to the forecast of e-commerce inventory demand. Inventory demand forecast has played an important role in supply chain management [2]. Therefore, it is of great significance to improve the accuracy of the Inventory demand forecast for organizing the inventory reasonably, reducing the logistics cost and improving the customer service level of e-commerce.

In the era of big data, data mining technology can be flexibly applied to forecasting inventory demand forecast based on sales data [3], which will greatly improve the accuracy of inventory demand forecasts and eventually achieve accurate quantity.

Scholars at home and abroad have made lots of attempt in demand forecasting in recent years. The methods can be roughly divided into time series models and machine learning algorithms. Traditional demand forecast methods mainly exploit time series analysis techniques [4]. The existing demand forecasting methods in the E-commerce domain have been largely influenced by state-of-the-art forecasting techniques from the exponential smoothing [5] and the ARIMA [6] families. Although time series models have been proven to be useful for demand forecasting, their forecasting ability is limited by their assumption of a linear behavior, which do not take external factors such as price changes and promotions into account [7].

There have been most of the studies using machine learning algorithms. Zhao K and Wang C propose a novel approach to learn effective features automatically from the structured data using the Convolutional Neural Network [8]. Bandara, Kasun, et al. attempt to incorporate sales demand patterns and cross-series information in a unified model by training a Long Short-Term Memory network (LSTM) that exploits the non-linear demand relationships available in an E-commerce

product assortment hierarchy [9].

According to the above literature review, a combination forecasting model based on Two-step Clustering and Random Forest is constructed to forecast demand in this paper.

The paper is organized as follows. In Section 2, the proposed model based on Two-Step Clustering and Random Forest is described in detail. In Section 3, an empirical analysis is made to verify the validity of the proposed forecasting model. In Section 4, the conclusions along with a note regarding future research directions are summarized.

2. The proposed model based on Two-Step Clustering and Random Forest

In this research, a combination forecasting model (short for the C-RF model) based on Two-step Clustering and Random Forest, which incorporates different clustering features into forecasting as influencing factors.

Step 1. The Two-Step Clustering Algorithm is applied to partitioning training set into *n* different clusters based on various features. Each cluster is denoted as

$$C_i, (i = 1, 2, \cdots, n) \tag{1}$$

Step 2. For each cluster C_i, the Random Forest Algorithm is used to train forecast models, denoted as

$$C_i - RF, (i = 1, 2, \cdots, n) \tag{2}$$

Step 3. Testing set is used to verify the feasibility of the trained $C_i - RF, (i = 1, 2, \cdots, n)$. In detail, the testing set is firstly partitioned into the corresponding cluster by the trained Two-Step Clustering model in Step 1. Then, the results of the testing set are forecast by using the corresponding $C_i - RF$ model.

Step 4. The single Random Forest model is compared with the $C_i - RF$ model. The evaluation indexes of the C-RF model and a single Random Forest model are calculated. Therefore, the optimal model can be determined.

3. Empirical analysis

In this paper, the experimental data are originated from data accumulation of an e-commerce enterprise, which are divided into the training set (from 1st day to 277th day), validation set (from 278th day to 347th day), and testing set (from 348th day to 381st day). The description of the experimental data is listed in Table 1.

Table 1. Description of the experiment data

Classification variables	Continuous variables
Commodity number; commodity name	Order quantity; item view; visitor number; order conversion rate; cumulative attention amount; added shopping cart quantity; shopping cart conversion rate; evaluation quantity; praise quantity; praise rate; flow ratio in the website; flow ratio out of the website.

Step 1. The Two-step Clustering algorithm in SPSS Modeler is used to cluster the experiment data.

Firstly, the “source” node in SPSS Modeler imports the data of training set 1 and sets the field format of the “type” node.

Then, the relevant parameters are set in the two-step clustering algorithm of “modeling” node, and the continuous field is standardized; the noise percentage of outlier processing is set to “25%”; and the distance measurement method is set to “logarithmic similar value”; the standard of

clustering is “BIC” (Bayesian Information Criterion).

Finally, the trained two-step clustering algorithm is obtained by running “flow”.

As illustrated in Figure 1, the training set of all SKU is partitioned into 4 clusters; the quality of clustering is“Good”; the size ratio is 1.93. It shows that clustering performs a good classification effect.

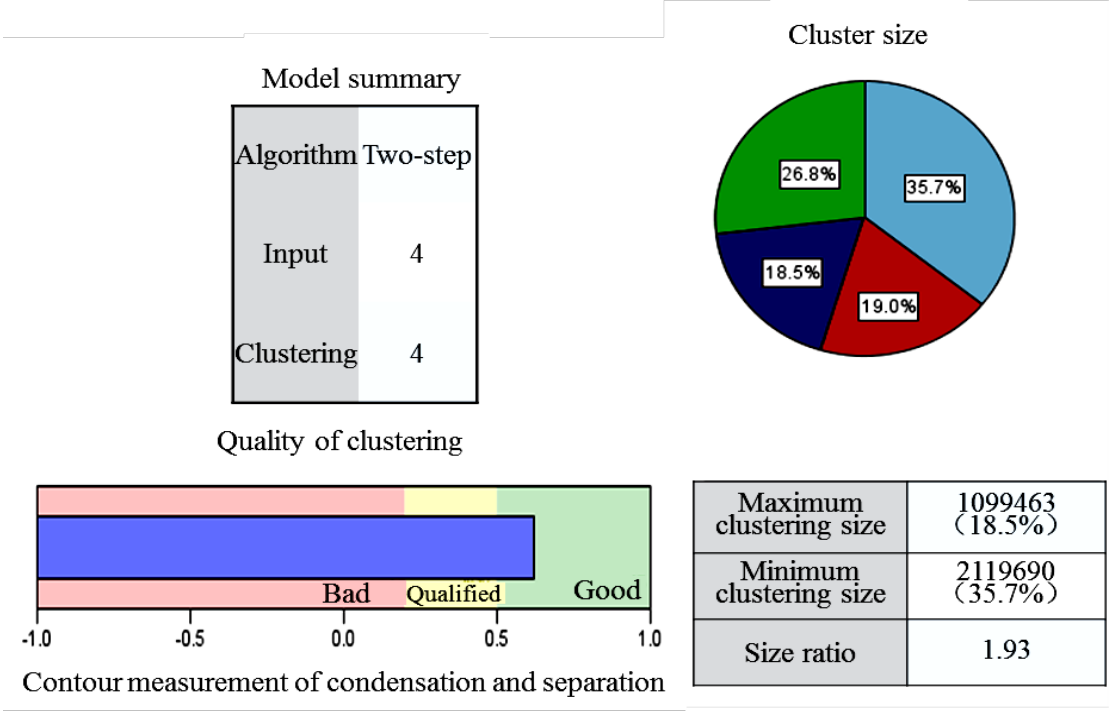

Figure 1. Model summary of the Two-step Clustering and clustering size

Step 2. For each cluster obtained in *Section 4.1*, Random Forest models based on Two-step Clustering are trained. Then, the validation set is used to check the built $C_i - RF$ models.

Step 3. The testing set is partitioned into the cluster C_4, so the corresponding $C_4 - RF$ model is used to calculate the forecasting results.

Step 4. The Random Forest model is trained with the training set. Then the validation set is used to check the built Random Forest model. Lastly, the results of the testing set are calculated.

Figure 2 shows the curve of actual values SKU_sales and two fitting curves of predicted values from the 348th day to the 381st day, which is obtained by the single Random Forest and the C-RF model. It can be seen that C-RF model has a better fitting performance to the original value than the single Random Forest model.

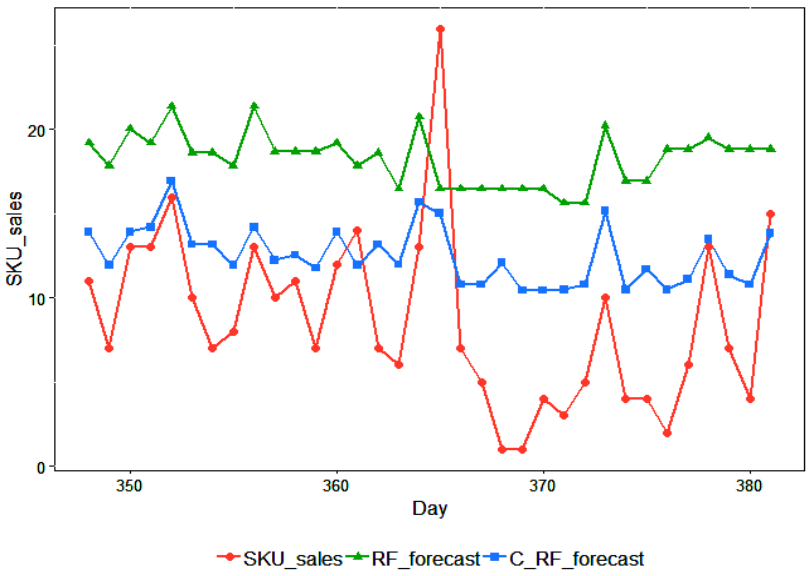

Figure 2. The comparison of C_RF model with the single Random Forest model

To further illustrate the superiority of the proposed C-RF model, the common evaluation indexes are listed in Table 2 for the C-RF model and the single Random Forest model. According to Table 2, it can be concluded that the superiority of the C-RF model is distinct compared with the single Random Forest, as its evaluation indexes are minimized.

Table 2. Evaluation indexes of the C-RF model and the single Random Forest model

Evaluation index	Training set		Validation set		Testing set	
	C4-RF forecast	RF forecast	C4-RF forecast	RF forecast	C4-RF forecast	RF forecast
Minimum error	-6	-79.852	-14.83	-20.1	-11.066	-16.805
Maximum error	429	232.766	30.159	25.105	10.983	9.521
Mean error	0.683	0.173	-5.758	-9.563	-3.844	-9.578
Mean absolute error	0.915	9.956	7.949	11.098	4.681	10.138
Standard deviation	3.005	26.058	7.32	7.57	3.972	4.667
Linear correlation	0.554	0.891	0.715	0.657	0.784	0.461
Occurence	445,366	277	70	70	34	34

4. Conclusion

In this research, a combination forecast model based on Two-step Clustering and Random Forest is proposed, which takes factors of demand into account. The Two-Step Clustering Algorithm is applied to partitioning data into different clusters. After that, the corresponding C-RF models are established for different clusters using the Random Forest.

To verify the effectiveness of the proposed C-RF model, the single Random Forest is employed for comparison. The experiment demonstrates that the C-RF outperforms the single model, and the clustering has improved the accuracy of the demand forecast. It is advisable for the e-commerce company to train different forecasting models for different commodities.

In future research, the forecast models combining clustering with other machine learning model are worth exploring and studying.

References

[1] Mcneely C L, Hahm J O. The Big (Data) Bang: Policy, Prospects, and Challenges [J]. Review of Policy Research, 2014, 31 (4): 304 – 310.

[2] Li T. Application of the Massive Data Accuracy Classification in E-commerce based on Big

Data [C]. Sichuan, Chengdu: Atlantis Press, 2015. 1227 - 1231.

[3] Lazcorreta E, Botella F, Fernández-Caballero A. Towards Personalized Recommendation by Two-Step Modified Apriori Data Mining Algorithm [J]. Expert Systems with Applications, 2008, 35 (3): 1422 - 1429.

[4] Wu J, Li Y, et al. Modeling a Combined Forecast Algorithm based on Sequence Patterns and Near Characteristics: An Application for Tourism Demand Forecasting [J]. Chaos, Solitons & Fractals, 2018, 108: 136 - 147.

[5] Gmbh S. Forecasting with Exponential Smoothing [J]. Springer, 2008, 26 (1): 204 - 205.

[6] Box G E P, Jenkins G M. Time Series Analysis, Forecasting, and Control [J]. Journal of the American Statistical Association, 1971, 134 (3).

[7] Zhang G P. Time Series Forecasting Using a Hybrid ARIMA and Neural Network Model [J]. Neurocomputing, 2003, 50 (none): 159 - 175.

[8] Zhao K, Wang C. Sales Forecast in E-commerce using Convolutional Neural Network [J]. 2017.

[9] Bandara K, Shi P, Bergmeir C, et al. Sales Demand Forecast in E-commerce using a Long Short-Term Memory Neural Network Methodology [J]. 2019.