ISBN 978-981-11-3671-9
Proceedings of 2017 the 7th International Workshop on Computer Science and Engineering (WCSE 2017)
Beijing, 25-27 June, 2017, pp. 465-469
doi: 10.18178/wcse.2017.06.080

Ticket price forecasting based on machine learning

Yuling Li [1 +,] Zhengmin Li [2] and Sujuan Qin [3]

[1] State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, 100876, China
[2] National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing, 100029, China
[3] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, 100093, China

Abstract. With the development of the aviation industry and the improvement of people's living standard, more and more people choose the aircraft as their way to travel, but the airline adjusts the price according to the revenue management in real time. Due to the large fluctuations in ticket prices, the price forecast has practical application value. This paper proposes a combination algorithm, combining the time series algorithm and random forests algorithm to model the ticket price data, and realizes the accurate prediction of the ticket price. The experimental results show that the combination algorithm model is more reliable by comparing the forecasting results with the actual results of each price model. The model is helpful for passengers to buy tickets and to save money.

Keywords: price forecasting, time series algorithm, random forests algorithm

1. Introduction

Corporations often use complex policies to vary product prices over time. The airline industry is one of the most sophisticated in its use of the revenue management in an attempt to maximize its revenue. The pattern of air ticket market varies from one country to another, depending on the quantity and structure of the supply and the demand. The aim of this research is to create a model, which could be used for airline tickets price forecasting for China cities. At present, there are many price forecast papers using data mining technology, but because the system of domestic airline ticket pricing discount is not mature enough.

In 2003, Hamlet algorithm is used to deal with the flight price, and the three kinds of algorithms are used to form a new model [1]. The method of data processing is used to mark the point process, and the process is predicted by using random forest and decision tree classifier [2]. A Lantseva, and K Mukhina establish a regression model to predict the price [3]. In 2015, W Groves and M Gini use feature selection techniques to give the lowest price for all flight forecasts [4]. However, because of different revenue management between China and other countries, foreign forecasting methods are not suitable for China's ticket price changes. The contribution of this work includes the following:

- Collection, processing and fusion of data from 9 cities in china;
- Analysis of macroscopic trends in price formation;
- The price forecasting model based on time series and random forest algorithm.

2. Brief Theoretical Background

In this paper, two different models have been used for forecasting. The first one is the time series model, whereas the second model involves random forest model.

+ Corresponding author. Tel.: +18813162153.
E-mail address: 18813162153@163.com.

In Ref. [5], they only used the time series model to predict the ticket prices. In our case, we use the time series model to predict the ticket prices and then employed random forest to optimize the errors. Before describing the complete methodology of our system, we would like to begin by presenting some theoretical background on these.

2.1. Time series

Time series analysis is a large and diverse subfield of statistics whose goal is to detect and predict trends. In this paper, time series model is constructed based on the first order moving average model [5]. To predict the price of other states in the same class in the time series model, an equivalence class is defined as follows: the set of states with the same flight number and departure time but different takeoff dates. Using the equivalence class, we predict that p_{t+1} will be:

$$\frac{\sum_{i=1}^{k} \partial(i) avg(p_{t-k+i})}{\sum_{i=1}^{k} \partial(i)} \tag{1}$$

Here, $\partial(i)$ is some increasing function of i; $avg(p_{t-k+i})$ is the average value of the price in the same equivalence class.

2.2. Random forest algorithm

As the name implies, it is to establish a forest in a random way, the forest which is made up of a lot of decision trees. The random forest algorithm can effectively reflect the interaction between the variables, taking into account the impact of various factors on fares.

3. Data Collection

We collected airfare data directly from a major travel web site. For the purpose of our pilot study, we have access to the domestic 18 routes, 9 cities ticket price data. We collect data 60 days in advance at twelve-hour intervals. The data we collected have the following characteristics: city of departure, destination, ticket purchase date, departure date, ticket options with the price, time of departure. Data processing: data cleansing, deleting data that is unrelated to the ticket forecast process.

4. Prediction Model

This paper constructs the ticket price forecasting model by using the combination of time series analysis and random forest. The model is constructed as follows:

- Data collection and preprocessing;
- Predict the price of tickets using time series algorithm ;
- Predict residual error using random forest algorithm ;
- Get ticket price forecast.

Using the model presented in this paper, we predict that p will be:

$$p_{predicted\ \ value} = p_{time} + RF_{error} \tag{2}$$

Here: $p_{predicted\ \ value}$ represents the final predicted price; p_{time} represents the predicted value of the time series; RF_{error} represents the predicted value of random forest.

4.1. System design

In this paper, we construct the time series-random forest combination model. We use the time series model to predict the ticket prices and then utilize random forest to optimize the errors. Time series-Random forest combination algorithm flow is shown in Fig.1:

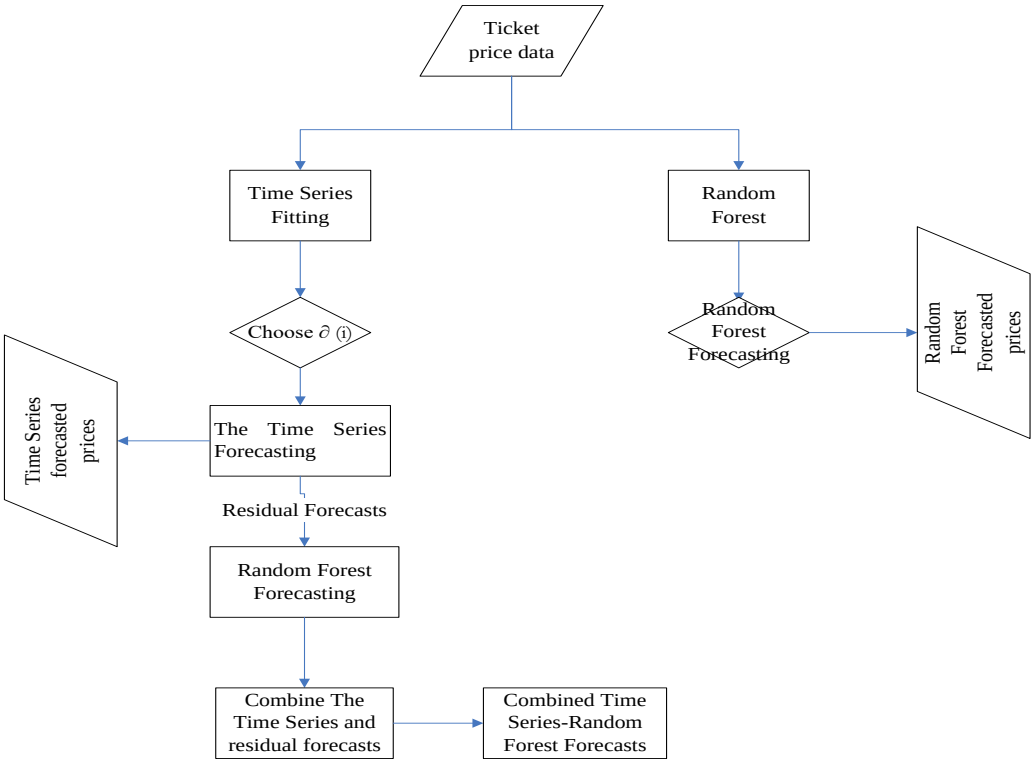

Fig.1: Complete system process, including time series and random forest algorithm

5. Experiments and Results

In this part, we carried out experiments using time series analysis, random forest algorithm, Time series-Artificial neural network and our combined Time series-Random forest. We use the flight route from PEK to HET to demonstrate the experimental results.

5.1. Data fitting results

5.1.1 The time series fitting

Based on the time interval, which is one of the most important factors for training according to the existing historical data, the time series algorithm is developed for our datasets. Taking the forecast of PEK to HET HU7175 flight as an example, the fitting results are shown in Fig.2. The fitting is an initial forecast and the errors are later corrected using the random forest models.

5.1.2 The random forest fitting

The random forest algorithm predicts the price of the ticket, and the results are shown in Fig.3. We can observe that the random forest can be a good predictor of the price trend.

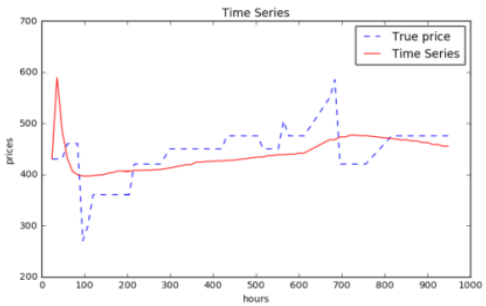

Fig.2: Time series - Fitted Data versus original flight

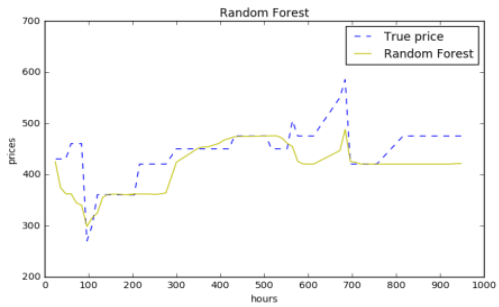

Fig.3: Random Forest - Fitted Data versus original for HU7175 for HU7175 flight

5.1.3 Time series-random forest fitting

The ticket price forecast results of the time series-random forest model are shown in Figure 4. Prioritize the important factors of the time interval, and then consider other factors as well as the interaction between

the factors. We can observe that the time series random forest combination model can fit well the ticket price.

5.2. Analysis of experimental results

Taking HU7175 flight as an example, given the date of departure, at different time intervals to predict the flight ticket price, and then compare the forecast results and the actual travel website published ticket prices. The experimental results are shown in Table 1. We can see that the prediction accuracy of our algorithm is higher than that of time series algorithm, random forest algorithm, and time series-artificial neural network algorithm.

Table I: Comparison Table of Forecast Price and Actual Price of Each Model

Time interval	Time series (price)	Random forest (price)	Time series-NN (price)	Time series-Random forest (price)	Actual price	Time series (error%)	Random forest (error %)	Time series-NN (error%)	Time series-Random forest(error%)
48/h	484.31	361.00	407.45	446.76	430.00	12.63	16.05	5.24	3.90
144/h	399.32	360.00	399.44	364.51	360.00	10.92	0	10.96	1.25
7/d	403.99	360.00	404.10	366.00	360.00	12.22	0	12.25	0
10/d	408.29	360.00	408.41	422.07	420.00	2.79	14.29	2.76	0.49
21/d	433.64	474.89	436.79	472.93	475.00	8.71	0.02	8.04	0.44

The resulting errors have been summarized in Table 2 and graphically displayed in Fig.5. The performances have been summarized in terms of the mean absolute percentage error (MAPE), mean square error (MSE), root mean square error (RMSE) and normalized root mean square error (NRMSE - percentage). Comparing the predictive results of each model, we can observe that the predictive effect of the Time series-Random forest combination model is best in the above models. The error of the model proposed in this paper is smaller than that of other models. But the running time of the model is relatively long.

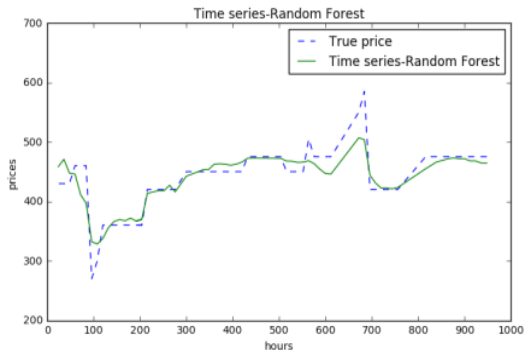

Fig.4: Time series - Random forest Fitted Data versus original for HU7175 flight

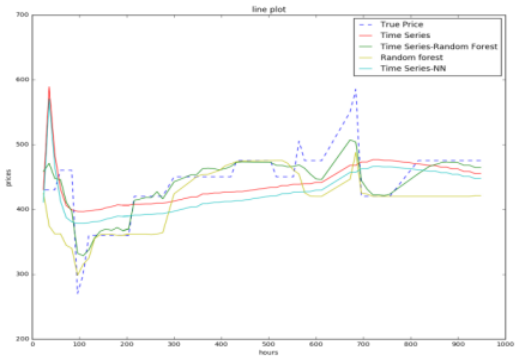

Fig.5: All models Fitted Data versus original for HU7175 flight

Table. II: Summary of the Results for HU7175

Algorithm	MSE	RMSE	NRMSE	MAPE	Run time (s)
Time Series	2259.54806257	47.5347037707	0.150903821494	9.031814052	13.668782
Random Forest	2191.92881875	46.8180394586	0.148628696694	7.58443622594	4.963284
Time series – Random forest	546.862525066	23.3850919405	0.0742383871127	3.59662290396	23.909368
Time series -ANN	2294.31565454	47.899015173	0.152060365628	9.19873759956	24.531276

6. Conclusion and Future Work

In conclusion, this paper presents a model of air ticket price forecast based on time series and random forest algorithm, in this experiment Time series algorithm, Random forest algorithm, Time series-Artificial neural network algorithm and combined Time series-Random forest models were used to predict ticket prices. The reliability of the model is proved by experiments. The construction of the new model provides strong support for passengers to purchase tickets.

In future, the method promoted here can be improved in running time. Create a method to improve the efficiency without changing the prediction accuracy.

7. Acknowledgements

This work is supported by NSFC (Grant No. 61502044), the Fundamental Research Funds for the Central Universities (Grant No. 2015RC23).

8. Reference

[1] Etzioni, Oren, et al. "To buy or not to buy: mining airfare data to minimize ticket purchase price." ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, Dc, Usa, August 2003:119-128.

[2] Wohlfarth, Till, et al. *"A Data-Mining Approach to Travel Price Forecasting."* International Conference on Machine Learning and Applications and Workshops IEEE, 2011:84-89.

[3] Gordiievych, Anastasiia, and I. Shubin. *"Forecasting of airfare prices using time series."* Information Technologies in Innovation Business Conference IEEE, 2015:68-71.

[4] Groves, William, and M. Gini. *"On Optimizing Airline Ticket Purchase Timing."* ACM Transactions on Intelligent Systems & Technology 7.1(2015):1-28.

[5] Gu Zhaojun, Wang Shuang, Zhao yi. *"A forecasting model of air ticket price based on time series. "* [J]. Journal of Civil Aviation University of China, 31.2(2013):80-84.

[6] Yaqub, Mohammad Umair, and M. S. Al-Ahmadi. *"Application of Combined ARMA-Neural Network Models to Predict Stock Prices."* The, Multidisciplinary International Social Networks Conference on Socialinformatics 2016, Data Science ACM, 2016:40.

[7] A Lantseva, K Mukhina, A Nikishova, S Ivanov, K Knyazkov,“Data-driven Modeling of Airlines Pricing”《Procedia Computer Science》, 2015, 66:267-276.