ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Shanghai, 19-21 June, 2020, pp. 57-61
doi: 10.18178/wcse.2020.06.010

Flight Delay Prediction Based on Elastic Neural Network

Chaoyang Lu, Shuze Hang [+] and Meize Dai

College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing, China

Abstract. Accurate prediction of flight delays is a difficult issue for airlines operation. Flight delays are affected by a variety of factors. It is difficult to accurately predict flight delay time from the perspective of traditional statistics. In order to reduce the data over-fitting, this study uses the genetic algorithm to select 21 related features. Then the regularized parameter L1 norm and L2 norm are introduced. Furthermore, the elastic neural network flight delay time prediction model is established to predict flight landing delay time. Finally, the accuracy within ±3 minutes tolerance is 83.954%, and the accuracy within ±5 minutes tolerance is 94.431%. The results show that the proposed model can improve the accuracy of flight delay prediction compared with the traditional simulation model.

Keywords: Air traffic, flight delay prediction, genetic algorithm, feature selection, elastic neural network, Machine learning

1. Introduction

Air traffic is often subject to delays due to environmental, weather and regulatory conditions. Domestic and foreign scholars have carried out an army of research on their predictions based on many uncertain factors of flight delay. The traditional method is mainly to establish the prediction model, including delay series prediction models such as time series, auto-regression, dynamic optimization and queuing theory[1][2]. Jinn[3] proposed the Cox prediction delay model, using the repeated chain effect to develop flight delay propagation, and assessing the level of flight delay and the impact on airline operations based on the risk level and the extent to which individual impact factors of flight delays affect airline operational reliability. However, this model considers few factors, and the assumption tends to be idealized, which cannot reflect the actual situation well; Subsequently, Robinson, D.P[4] proposed a delay propagation model, which calculated the possibility of delay based on a large number of passenger trip data and flight data, but the prediction accuracy is not high; Wang P T R[5]established a recursive model of flight delay propagation, optimizing delay propagation equations by separating variables and immutable variables that affect flight delays.

Because There are many potential factors affecting flight delays, the mathematical model based on certain assumptions and ignoring some conditions have many limitations in flight delay prediction and it is difficult to fully consider all the influencing factors[6][7]. In recent years, with the rise of big data, some scholars[8][9][10] began to predict flight delays through data mining to solve these problems. Álvaro[11] et al. studied Markov chain and Bayesian network model to predict delay time and obtain reliable accuracy in busy airports in Europe; Kim YJ [12]used cyclic neural networks to study the delay state and applied the model to Atlanta airports and other airports with good generalization and accuracy. This paper is mainly divided into the following parts: firstly analyse the difficulty of flight delay prediction, then select important features based on random forests, and finally propose a delay time prediction model based on elastic neural network.

[+] Corresponding author. Tel.: +86 13862221066
E-mail address: declanvala96@163.com.

2. Analysis of the Difficulty of Flight Delay Prediction

Flight operation is subject to many factors such as the controller's command, scene conditions, weather and so on, it is easy to cause flight delays. The main difficulty in studying the delay of flight is that a large part of operation data of flight is difficult to collect and many factors cannot be considered, which leads to the low accuracy of some analogy methods.

The sample of this study comes from the 2017 flight data of the six major airports announced by the US National Traffic Statistics Bureau, totaling 1.2 million flights. Firstly, we find the "abnormal point" of delay time according to the 3σ principle, and use the nuclear density estimation to plot the delay time nuclear density of the six airports, as shown in Figure 1. On this basis, we draw the outliers and non-outliers of the six major airports, as shown in Figure 2.

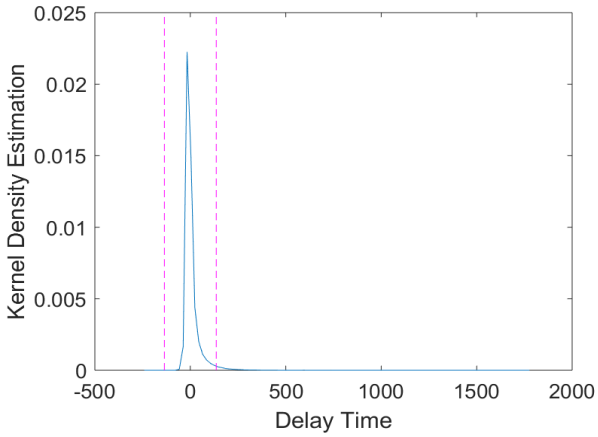

Fig. 1: Estimation of delay time density at six major airports

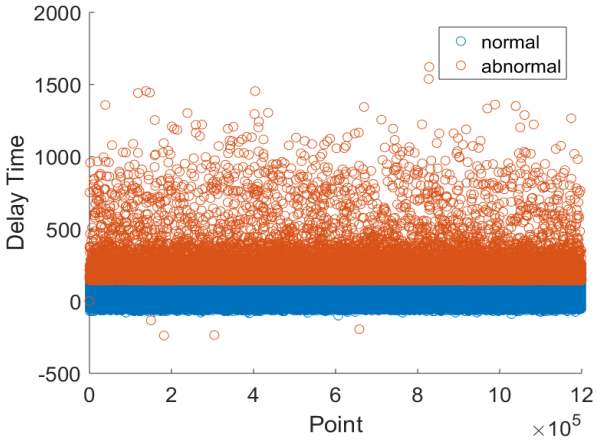

Fig. 2: Six major airports delay time distribution

In Fig.1, the part enclosed by the red dotted line is the probability distribution of normal value, the abnormal part is the red part, it can be found that the current flight delay distribution of the six major airports is concentrated between 0 and 500. In Fig. 2, the sample mean is 5.52, and the variance is 45.31. According to the 3σ principle, Fig. 2 plots 21182 abnormal values. It can be concluded that the distribution of flight delay time does not conform to normal distribution and Poisson distribution, and is an irregular distribution, which makes it difficult to accurately predict flight delay time.

In this paper, we use K-means clustering method to cluster delay time. According to the principle of maximum expectation, five categories are classified. The delay time clustering graph shown in Figure 3 is obtained:

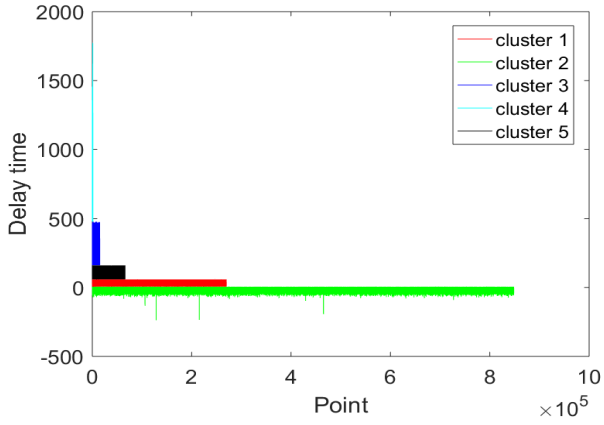

Fig. 3: Six major airports delay time clustering distribution

By analyzing the delay time obtained by clustering results, we find the fourth category of delay time variance is 199, but the number of samples is only 1139, which may be due to extreme weather, air control conditions and so on. The large time span and the unbalanced distribution of the categories also make it more difficult to predict the delay time.

According to the above analysis, the main difficulty of flight delay prediction is that the accuracy of prediction is not enough due to the irregular distribution of flight delay time. This paper will solve this problem by machine learning method based on historical data of flight operation.

3. Feature Selection Based on Bayesian Genetic Algorithms

The standardized data of the 2017 annual flight delay operation of The National Traffic Statistics Bureau is selected as a sample set, which is 1,231,359 in total. The 28 features includes: the flight date, estimated arriving time, wheels off time, wheels on time, destination airport ID, the distance between the departure and the destination, the group of the initial variable will be built as: $X'=\{AT,\mathrm{WO},\mathrm{OCMID},...,DIST\}$ $feature=28$.

Table 1 shows a part of the sample data, five samples and six groups of features will be randomly selected, and the features will seperately correspond to: the identify number assigned by American DOT, the flight number, the origin airport state ID, the city market ID, the estimated time of departure delay and the sliding time.

Table 1: Units for Magnetic Properties

	DOTIDRA	FNRA	OASID	OCMID	DD	TO
1	20366	5571	1039705	30397	-8.0	168.0
2	20304	2966	1393004	30977	112.0	165.0
3	19790	2238	1039705	30397	-3.0	141.0
4	20366	3946	1393004	30977	-7.0	139.0
5	19930	101	1474703	30559	-7.0	137.0

According to the method of the feature selection of the genetic algorithm, select the features of the initial given 28 feature $X'=\{AT,\mathrm{WO},\mathrm{DSF},\mathrm{OCMID},...,DIST\}$, $feature=28$ substitute the test samples into the model for iteration, and exit the circulation when the fitness reaches 0.95.

The new feature group includes 21 features, compared with the original group, 7 unrelated features are removed, including: DDG,OW,OSF,CANCEL,DIVER,FLIGHTS,YEAR, which are separately correspond to： Distance interval of flight segment, origin & destination airportID, intervals of departure delay（15 minutes will be recorded as 1）,the world time of origin airport, the origin state ID, whether the pre-order flight is cancelled, whether the aircraft is transferred（1 means yes）,Number of flights at the same time, years.

On the basis of selecting the 21 remaining features, the delay prediction model will be built based on the elastic neural network in Chapter 4, and a prediction of flight delay of six major airports of the United States will be made according to real cases.

4. Delay Time Prediction Model Based on Elastic Neural Network

4.1. Delay Time Prediction Model

In order to reduce the occurrence of over-fitting phenomenon, the L1 norm and L2 norm are introduced as the a priori regular training regression model. Finally, the flight delay prediction model based on elastic network is constructed.

The elastic neural network controls the sparsity in the regression process through the L1 and L2 norms, and finally can obtain a model with only a few parameters being non-zero sparse. In the case of such a large sample size, the elastic neural network reduces the time cost of training with an exceptionally fast convergence.

According to the elastic neural network, L1 and L2 norms are introduced to regularize them, and the final cost function is:

$$J\cos t = \frac{1}{2m}\sum_{i=1}^{m}\left(y_i - f(x_i)\right)^2 + \alpha\lambda\|w\|_1 + \frac{\alpha(1-\lambda)}{2}\|w\|_2^2 \tag{1}$$

Where, α is disciplinary factor, the larger the value is, the more sparse the data set will be; λ is norm weight, The second term is L1 norm and the third term is L2 norm. The minimum cost function is required in the iterative process. According to the factor terms of L1 and L2 regularization, the feature with low contribution rate to the sample will be eliminated. This paper adopts the coordinate descent method to solve the cost function.

Coordinate descent is a non-gradient optimization algorithm. The local minimum of the cost function is found by one-dimensional search along a coordinate direction in the iterative process. The main principle is to optimize the target to iteratively reduce the loss function on the n axes of w. When w_i (i = 1,2,…,21)on all axes converges, *J* cost reaches the minimum. At this time, *w* is the optimal result. Take an initial value for *w*, denoted as $w^{(0)}$, and the number in parentheses represents the number of iterations. For the iteration of round *k*, starting from $w_i^{(k)}$, $w_i^{(k)}$ was successively calculated, and the process of $w_i^{(k)}$ iteration is as follows:

$$\begin{aligned}
&w_1^{(k)} \in \arg\min_{w_1}\left(w_1, w_2^{(k-1)}, \ldots, w_n^{(k-1)}\right)\\
&w_2^{(k)} \in \arg\min_{w_2}\left(w_1^{(k)}, \ldots, w_{i-1}^{(k)}, w_i, w_{i+1}^{(k-1)}, \ldots, w_n^{(k)}\right)\\
&\ldots\\
&w_n^{(k)} \in \arg\min_{w_n}\left(w_1^{(k)}, w_2^{(k)}, \ldots, w_n\right)
\end{aligned} \tag{2}$$

It can be seen from the above formula, $w_i^{(k)}$ is the w_i that minimizes J cost, In this case, only $w_i^{(k)}$ is the variable, the rest are constants, so the minimum is easily obtained by derivation.

4.2. Flight Delay Time Prediction Analysis

The flight delay prediction model is divided into three parts. The first step is to preprocess the original data; The second step, the genetic algorithm is used to extract the features of the original data set, and the variables that can contain 95% of the sample features are extracted; The third step is to build an elastic neural network model to train the data set and finally achieve the goal of predicting flight delay time.

Before predicting delay time, it is necessary to determine the correctional factor α and λ norm weight in equation 1. Rule of thumb is to assign α to 1, λ is marked with 20 equal marks in [0, 1]. Use the grid search method to adjust the model, and adjust the reference to draw the ROC curve (accuracy is the tolerance within ±3 minutes), As shown in Figure. 4 the x-coordinate shall be the value of λ, and the y-coordinate shall be the accuracy:

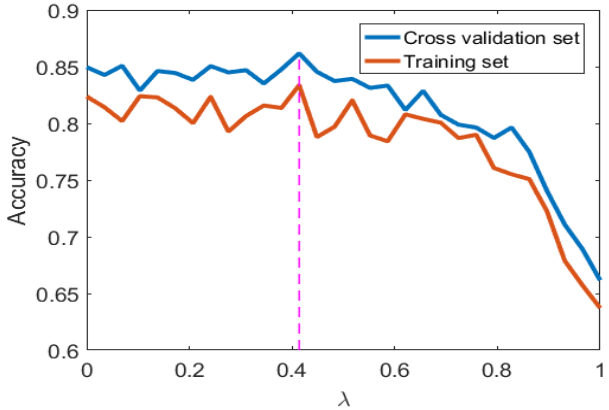

Fig. 4: Regularization parameter λ curve

Since the sample delay time is only an integer, this paper rounds up the predicted sample and calculates its accuracy. The blue curve is the accuracy of the cross validation set, and the red curve is the accuracy of the test set. In the λ curve drawn by the grid search method, it can be found that when the value of λ is 0.433, the prediction accuracy of the test set delay time reaches 83.332%, and the prediction accuracy of the cross-

validation delay time reaches 86.171%, which basically realizes the flight delay. Accurate prediction of time. It was finally determined that $\lambda = 0.433$ was the best parameter.

Similarly, take $\lambda=0.433$ and take 30 equal parts in the [-5, 1] interval according to experience α. Using the grid search method, the model is adjusted, and the α curve is adjusted (the accuracy is within ±3 minutes), the highest accuracy is determined when α is -2.517, and the selection of this parameter ends.

5. Conclusions and Discussion

In this paper, by plotting the delay time distribution map and the delay analysis, it is found that the flight delay time distribution is uneven and difficult to predict. In order to solve this problem, a genetic algorithm feature selection model based on Bayesian is established. and the original 28 flight operation characteristics were filtered to 21 features using the flight data of the US National Traffic Statistics Bureau. On this basis, the regularization parameter L1 norm and L2 norm are used to establish the elastic network flight delay time prediction model, and then the model is trained and tested with the 2017 and 2018 delay data. The results show that the delay time prediction accuracy within ±3 minutes tolerance reaches 83.954%, the accuracy rate within the 5-minute tolerance reached 94.431%. The model is finally verified by the decision coefficient and interpretation variance.

6. References

[1] Modelling delay propagation within an airport network[J]. Nikolas Pyrgiotis,Kerry M. Malone,Amedeo Odoni. Transportation Research Part C. 2011

[2] Estimating Airport System De-lay erformance. Jerry D Welch,Richard TLloyd. 4th USA/Europe Air Traffic Management R&D Seminar. 2001

[3] Jinn-Tsai Wong,Shy-Chang Tsai. A survival model for flight delay propagation[J]. Journal of Air Transport Management,2012,23.

[4] Robinson D P, Murphy D J. Enhanced flight delay data for ASQP carriers[C]// Integrated Communications, Navigation & Surveillance Conference. 2012.

[5] Wang P T R, Schaefer L A, Wojcik A L A. Flight Connections and Their Impacts on Delay Propagation[C]// Digital Avionics Systems Conference. IEEE, 2003.

[6] Peterson E B, Neels K, Barczi N, et al. The Economic Cost of Airline Flight Delay[J]. Journal of Transport Economics and Policy (JTEP), 2013, 47(1):-.

[7] Robust Airline Schedule Planning: Minimizing Propagated Delay in an Integrated Routing and Crewing Framework[J]. Dunbar, Michelle,Froyland, Gary,Wu, Cheng-Lung. Transportation Science. 2012 (2)

[8] Liang W, Li Y. Research on optimization of flight scheduling problem based on the combination of ant colony optimization and genetic algorithm[C]// IEEE International Conference on Software Engineering & Service Science. IEEE, 2014.

[9] Geng, Xi. [IEEE 2013 5th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC) - Hangzhou, China (2013.08.26-2013.08.27)] 2013 5th International Conference on Intelligent Human-Machine Systems and Cybernetics - Analysis and Countermeasures to Flight Delay Based on Statistical Data[J]. 2013:535-537.

[10] Rebollo J J, Balakrishnan H. Characterization and prediction of air traffic delays[J]. Transportation Research Part C: Emerging Technologies, 2014, 44:231-241.

[11] Álvaro Rodríguez-Sanz,Fernando Gómez Comendador,Rosa Arnaldo Valdés,Javier Pérez-Castán,Rocío Barragán Montes,Sergio Cámara Serrano. Assessment of airport arrival congestion and delay: Prediction and reliability[J]. Transportation Research Part C,2019,98.

[12] Kim Y J, Choi S, Briceno S, et al. A deep learning approach to flight delay prediction[C]// 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC). IEEE, 2016.