Flight Delay Prediction Based on Characteristics of Aviation Network

Tan Zhou¹, Qiang Gao¹, Xin Chen² and Zongwei Xun¹
¹ Civil Aviation College, Nanjing University of Aeronautics and Astronautics, 211106, China
² Management Science and Engineering College, Nanjing University of Finance and Economics, 210046, China

Abstract. In recent years, the increasingly serious flight delay affects the development of the civil aviation. It is meaningful to establish an effective model for predicating delay to help airlines take responsive measures. In this study, we collect three years’ operation data of a domestic airline company. To analyse the temporal pattern of the Aviation Network (AN), we obtain a time series of topological statistics through sliding the temporal AN with an hourly time window. In addition, we use K-means clustering algorithm to analyse the busy level of airports, which makes the airport property value more precise. Finally, we add delay property and use CHAID decision tree algorithm to train the data of an airline for nearly 3 years and use the training model to predict recent half a year delay. The experimental results show that the accuracy of the model is close to 80%.

1 Introduction

With the developing of civil aviation, the number of airports and flights are increasing sharply. As a result, flight delay gets so serious that limits the development of airports and airlines.

In the perspective of aviation network, Nikolas Pyrgiotis¹ proposed a phenomena of propagated delay in large-scale aviation network (AN). In addition, Guimera², Jianhong Mou³, Liu⁴, Xiaozhou Zeng⁵ et al. conducted an analysis of modeling and topology of aviation network based on complex network theory. In terms of delay prediction, decision tree⁶ is a kind of data mining model which builds a classification model through information entropy in an unordered and irregular data set. At present, decision tree models are applied in many areas, such as C4.5 for network traffic prediction⁷, financial warning⁸, remote sensing image classification⁹, and Chinese character recognition¹⁰, etc. And CHAID is used for credit card risk management¹¹ and personal income forecast¹². In this paper, combined with the dynamic topology characteristics of aviation network, a prediction model based on CHAID decision tree is proposed. The study is conducted from the perspective of the airline and is tested with three years’ operation data of a major domestic airline.

2 Temporal Characteristics of Topology

2.1 Sources of data

We collect actual flight operation data of a major domestic airline from July 1, 2013 to July 1, 2016, a total of 2,114,122 pieces of data. Among them, there are 1,421,359 delayed flights, accounting for 67%, and total arrival delay is 125,476,116 minutes.

Each flight information includes flight date, flight number, aircraft type, agent, nature of flight, actual takeoff station, actual landing station, planned takeoff station, planned landing station, actual/planned departure time, actual/planned arrival time, flight time, takeoff delay etc.

Extract four important segments as training set.

| Training set      | Number |
|-------------------|--------|
| Training set 1 PEK-SHA | 18321 |
| Training set 2 SHA-CAN  | 12061 |
| Training set 3 PVG-XIY | 8103  |
| Training set 4 KMG-PEK | 9801  |

2.2 Static characteristics of topology

(1) The topology of the network is shown in Figure 1. The average shortest path length <l> = 2.412, the average clustering coefficient <C> = 0.663, and the network diameter <d> is 6, indicating that the separation between the two airports is very small, and the average can be reached in as few as 2 transfers. It needs no more than four connections for any two farthest airport to reach each other. The network has a shorter path length and a higher clustering coefficient, which shows typical small-world network characteristics. Therefore, the flight delay of any node will be quickly propagated to other nodes.

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).
2. Dynamic Characteristic of Aviation Network

Time Window. We analyze the temporal characteristics of AN by sliding time window by one hour.

In this method, we use a contact sequence of quadruplets \((i, j, t'_i, t'_j)\) to represent aviation network, where \(i, j\) denote airports and \((t'_i, t'_j)\) are the takeoff and landing times of the flight. The edges between nodes appear at time \(t'_i = \{t'_i(1), t'_i(2), \ldots t'_i(n)\}\), which is ordered as \(t'_i(a) < t'_i(b)\). Since we assume that edges are established when the flights start to end, edges maybe overlapped because of the duration. We divide the daily AN into consecutive sub-networks with each network being constructed of flights with \(e_{ij}(t) = 1\).

\[
e_{ij}(t) = \begin{cases} 1, & t_i \leq t \leq t'_i \\ 0, & \text{otherwise} \end{cases}
\]

The connections contain departure, arrival and ongoing flights. The time window is shown as:

\[
\begin{align*}
T_w(m) &< t_i \leq T_w(m) + \Delta t \\
T_w(m) &< t'_j \leq T_w(m) + \Delta t \\
t_i < t_j &< T_w(m) + \Delta t \\
t_i < t_r &> T_w(m) + \Delta t
\end{align*}
\]

Where \(T_w(m)\) represents the initial time of the mth time window; \(\Delta t\) denotes the length of time window, and is set as \(\Delta t = 1\ h\) in our study.

Change in the number of flights. We start analyzing the temporal traveling pattern by calculating the change in the number of flights \(N_f\). As shown in Figure 3. The horizontal axis is time period, and the vertical axis is the number of flights. When the number is at a high level, it is easy to make airports out of order and the delay happens. The number starts from zero at 6:00 and increases dramatically during the following 3 hours. It maintains high level from 10:00 to 17:00, then there is a relatively slight decrease and the number returns to zero from 0:00 to 6:00.

3 Airport clustering analysis based on topological features

Observing the above training sets, we found delay of different airports also has a large gap. It can be speculated that the busyness of the airport has some effect on delay. The statistics of the original data were collected to obtain an average delay of 303 airports. In addition to the airports without delay, the effective data is 198 groups.

4 Conclusion: By observing Figure 4, it is found that there is a large gap between the delays at different airports. Although the busy airports have a high service rate, the traffic exceeds its capacity. Thus the operating status is in a congested state and is prone to delays. Therefore, it is necessary to consider the extent of the
airport's busyness as an attribute that influences flight delay. 
(1) Study the factors that influence airport delay and create the cluster index system. 
Based on the topological characteristics analyzed in Chapter 2: 
1. Shorter average path length and higher clustering coefficient can both lead to higher delays, because flight delay of any node will be quickly propagated to other nodes. 
2. The higher the average degree is, the higher the delay would be, because delay at the hub airports can be more likely to transmit to a wide range of airport nodes. We add in some social characteristics of the airports and build a clustering index system.

Table 3: Airport clustering index system

| Num | Influencing factors | Num | Influencing factors |
|-----|---------------------|-----|---------------------|
| 1   | Path length         | 5   | GDP                 |
| 2   | Clustering coefficient | 6    | Total population    |
| 3   | Degree              | 7   | Takeoffs and landings |
| 4   | Airport throughput capacity | | |

(2) Select airports with severe delay and establish airport samples

![Figure 5. The number and sum of delay](image)

Since the training set in this paper is selected from first-line airports, if all airports are used for classification, these airports will be difficult to divide. Taking the difficulty of data acquisition into account, the top ten airports with high delay are selected for clustering. The result is shown in Table 4.

Table 4. Airport clustering result

| Airport | Clustering | | | | Airport | Clustering |
|---------|------------|---|---|---|---------|------------|
| PVG     | 1          | NKG | 3 |
| SHA     | 2          | CAN| 2 |
| KMG     | 2          | TAO| 3 |
| PEK     | 1          | CTU| 2 |
| XIY     | 3          | WUH| 3 |

After calculation, Sig. of the cluster analysis is not higher than the significance level of 0.05, so it can be considered that there are significant differences among the above 7 variables. The clustering is ideal, and then we can use the airport cluster for delay forecasts at different airports.

4 Flight delay prediction based on decision tree

4.1 Attribute selection and division

In the flight forecasting problem, we need to filter attributes that are related to flight delay to ensure good forecasting results. We filter the attributes of the original data to obtain 9 available attributes. The specific attributes are shown in Table 5. The following describes the classification basis and classification results.

Table 5. Data set property description

| Number | Attribute         | Description |
|--------|-------------------|-------------|
| 1      | Solarterm         | The month of the flight date, an integer in the range of 1-12 |
| 2      | PlancType         | Aircraft types, classified into 3 categories according to the number of seats, and valued as 1, 2, and 3. |
| 3      | DepTimeofPlan     | Planned departure time, valued as 1, 2, 3, 4 |
| 4      | ArrTimeofPlan     | Planned arrival time, valued as 1, 2, 3, 4 |
| 5      | Origin            | Original airport, according to the result of clustering, the values are 1, 2, 3 |
| 6      | Destination       | Destination airport, according to the result of clustering, the values are 1, 2, 3 |
| 7      | FlightTask        | Flight mission attributes, such as regular shift, overtime, charter, etc. |
| 8      | FlightType        | Flight attribute, the values are 5 kinds of attribute codes, such as international flight J, domestic flight N, |
| 9      | Delay             | Delay level, valued as 1, 2, 3, 4 |

(1) Classification of flight dates
It is divided by month, divided into 1-12 months. 
(2) Classification of aircraft type
Aircraft types are divided according to the number of seats, as shown in Table 6.

Table 6. Classification of aircraft type

| Aircraft types | Identification | Number | Proportion |
|----------------|----------------|--------|------------|
| A320 B737      | 1              | 3064   | 15.8%      |
| B757 B767 B777 | 2              | 2833   | 14.8%      |
| A330 33E       | 3              | 13542  | 69.4%      |

(3)&(4) Classification of planned departure/arrival time
The number of flights at the airport during the time slot determines the airport's peak hours. Referring to the analysis in section 2.3, we divide the time period as Table 7.

Table 7. Classification of scheduled departure/arrival time

| Time period | Identification | Number | Proportion(%) |
|-------------|----------------|--------|---------------|
| 0:00—6:00   | 1              | 4450   | 22.9          |
| 6:00-10:00  | 2              | 5276   | 27.1          |
| 10:00-17:00 | 3              | 6556   | 33.7          |
4.2 Experimental results

Experiment 1 Analysis of a single segment. The training set I PEK-SHA is trained using CHAID decision tree, and the decision tree is obtained as shown in Figure 6 and used to test the data from July 1, 2016 to January 2017 to obtain the confusion matrix as Table 9.

| delay(min) | identification |
|-----------|----------------|
| <25       | 1              |
| 25-60     | 2              |
| 60-120    | 3              |
| >120      | 4              |

Table 9. Confusion matrix of training set 1

![Figure 6. Decision tree of training set 1](image)

Table 11. Confusion matrix of mixed training set

| Classification | Predicted |
|----------------|-----------|
| Observed       | 1.00      | 2.00      | 3.00      | 4.00      | Correct percentage |
| 1.00           | 10883     | 1323      | 235       | 871       | 81.8%              |
| 2.00           | 2095      | 4415      | 275       | 367       | 61.7%              |
| 3.00           | 1077      | 943       | 1324      | 152       | 37.9%              |
| 4.00           | 56        | 236       | 16        | 6006      | 95.1%              |
| Total          | 59.8%     | 12.9%     | 2.8%      | 24.4%     | 84.7%              |

4.3 Comprehensive analysis

The six sets in Experiment 1 were to predict a single segment. "Flight Date", "Aircraft type", "Planned Arrival Time", "Planned Departure Time", "Mission attribute", "Departure airport level", and "arrival airport level" are taken into account. The attributes "Flight Mission Attribute" and "Flight Attributes" are abandoned. Both properties are invalid. The average accuracy of forecast results reached 79.54%, close to 80%.

Experiment 2 is to predict multiple segments of a single training set. The six attributes of "flight date", "model", "planned arrival time", "planned departure time", "mission attribute", "departure airport level", and "arrival airport level" are taken into account. In the result, six properties are reserved. It is effective to analyze the network topology and add its characteristics into clustering aircrafts before prediction.
Concluded from the above seven groups of experiments, we can find that when the delay level is 1 and 4, the correct rate of the decision tree is relatively high, but the correct rate is low in the 2 and 3 categories. The reason may be related to the tilt of the training set. Decision-making path is mutually exclusive, and most of the data in the training set are at 1, 4 and it is speculated that over-fitting may occur. For this problem, pruning can be used to reduce the influence of over-fitting.

In the previous study, we divided the flight delays into four levels, which are represented by 1-4. The greater the number is, the more severe the delay. If we only consider two conditions: delay and not delay, although the accuracy rate may decrease, the correct rate of prediction should increase. According to the 2016 civil aviation flight normal statistics method (the definition of flight delay refers to the condition where actual arrival time exceeds scheduled arrival time more than 15 minutes). We change the flight attributes to 0 and 1, 0 means no delay, 1 denotes delay and then obtain the correct rate of four training sets, as shown in Figure 7.

Figure 7. Comparison of the two and four types of delay attributes

In the experiment, it is found that more detailed delay attribute partitioning will lead to a decrease in the precision. The more classification values are, the more difficult to classify. For the value of delay level, the degree of subdivision should be increased as much as possible while ensuring a certain accuracy rate. If the data is skewed to a large extent, for example, 2/3 data is concentrated in <25 minutes (level 1), the level 1 should be more subdivided, and the delay of >25 should be more roughly classified.

5 Conclusion

From the perspective of the airline, this paper constructs a delay tree prediction model for forecasting delay. Firstly, we analyze the topology characteristics of aviation network, and then combine node attributes with K-means clustering algorithm to classify the busy level of airports, which improves the timeliness of the classification and the accuracy of the prediction. Finally, we use three years’ operational data of a major domestic airline network to conduct experiments. The results show that the accuracy rate of the established model is close to 80%, which is of high prediction accuracy. The insufficiency of this paper lies in the tilt characteristic of the data, which is too concentrated in the delay level 1 and 4, which leads to low accuracy of the 2 and 3 levels of prediction. Our future research will focus on resolving this phenomenon of over-fitting. There are two methods to be adopted. One is to prune decision trees, the other is to classify the data sets more precisely in dense data level, and roughly divide in loose areas to solve the problem of over-fitting.

References

[1] Pyrgiotis N., Malone K M., Odoni A. Modeling Delay Propagation within an Airport Network. Transportation Research Part C: Emerging Technologies, 27: 60-75(2013).
[2] Uimera R., Amaral LAN. Modeling the Worldwide Airport Network. The European Physical Journal B-Condensed Matter and Complex Systems, 38 (2): 381-385(2014).
[3] Xiaozhou Zeng. Analysis of China’s Aviation Network Structure Based on Complex Network Theory. Nanjing: Nanjing University of Aeronautics and Astronautics (2012).
[4] Liu Hon KUN. Empirical study of Chinese city airline network. Acta Physica Sinica-Chinese Editin-56,106-112(2007).
[5] Quan Shao, Yan Zhu. Analysis of Flight Delay Propagation Based on Complex Network Theory Aeronautical Computing Technique (2015).
[6] Han J, Kamber M, Pei J. Data mining: Concepts and techniques (2006).
[7] XU Peng, LIN Sen. Internet traffic classification using C4.5 decision tree. Journal of Software, 20(10):2692-2704(2009).
[8] ZHAO Jing Xian. A model of financial distress early-warning based on decision tree. Journal of Harbin University of Commerce: Social Science Edition, 101(4):97-99(2008).
[9] YUAN Lin Shan, Du Pei Jun, Zhang Hua Peng, et.al. CBERS imagery classification based on decision tree and performance analysis[J]. Remote Sensing for Land & Resources, 2:92-98(2008).
[10] JIN Wen,YUAN Chun-fa. Identification of Chinese Unknown Word Based on Decision Tree. Journal Of Chinese Information Processing,18, 2 (2008).
[11] PENG Shi-feng. Design of Credit Card Risk Management System on CHAID Decision Tree. Shanghai: FU Dan University (2013).
[12] HUANG Qi. Analysis of Personal Income Based on CHAID Decision Tree. Mathematical Theory and Applications, 29, 4(2009).
[13] Zhang Wen-Tong, Zhong Yun-Fei. IBM SPSS data analysis and mining real case. Tsinghua University Press, (2013).