Predicting flight delay based on multiple linear regression

Predicting flight delay based on multiple linear regression

Yi Ding*

* School of Computer, Wuhan Vocational College of Software and Engineering, Wuhan, 430205, China

Abstract. Delay of flight has been regarded as one of the toughest difficulties in aviation control. How to establish an effective model to handle the delay prediction problem is a significant work. To solve the problem that the flight delay is difficult to predict, this study proposes a method to model the arriving flights and a multiple linear regression algorithm to predict delay, comparing with Naive-Bayes and C4.5 approach. Experiments based on a realistic dataset of domestic airports show that the accuracy of the proposed model approximates 80%, which is further improved than the Naive-Bayes and C4.5 approach approaches. The result testing shows that this method is convenient for calculation, and also can predict the flight delays effectively. It can provide decision basis for airport authorities.

1 Introduction

With the rapid development of the national economy, the demand for air transport has increased dramatically. The flight delay has become more and more serious, which directly causes serious damage to the image of civil aviation services. For passengers, flight delay caused the inconvenience of travel, bad mood, as well as the double loss of time and economy; for the airport, the delay of the flight seriously affects the normal operation of the airport; for airline, frequent flight delay not only bring huge economic losses to the airline, but also affect the reputation of the airline. Flight delay has become the shackles of the development of the aviation industry.

Today, flight delay has not only become a problem for the majority of travelers, but also the world's civil aviation industry problems. In order to solve the problem of flight delay, Civil Aviation Administration of China has developed a lot of programs to reduce the average flight delay time, to enhance the efficiency of flight operations. However, flight delay was caused by many reasons. The main factor is that the capacity of airport and airspace is insufficient. For other reasons, such as the weather, the airport scheduling, the company plan, passengers and luggage etc. may cause flight delay. There is a chain reaction of the flight delay. When flight delay occur, if the plan is compact, it will affect the take-off or landing of next flight on time, thereby indirectly affect more downstream flights and airports. If we could timely forecast the flight delay, we can take necessary measures to reduce the economic and credit losses due to it. Obviously, it is very important and necessary to predict the flight delay in real time.

In this paper we established a multiple linear regression model using departure delay and route distance to predict arrival delay, and presented the design and implementation of a flight delay prediction system. The researching value of this paper can be listed as follows: 1) Improving flight delay prediction accuracy and the decision-making ability of the relevant departments; 2) Promoting the application of computer simulation technology in civil aviation; 3) Serving for the construction of civil aviation information; 4) Making technology accumulation for flight delay early warning and emergency decision support system. The rest of this paper is organized as follows: In section 2, we

provide a brief review towards some related work on flight delay. Data used in this study are presented in section 3. Our method is presented in Section 4. The design and implement of system is described in Section 5. Our detailed Experimental results are presented in section 6. Conclusion and future work is presented in Section 7.

2 Related work

Flight delay resulted in significant costs to airlines, passengers and society. Such high delay costs motivate the analysis and prediction of air traffic delay, and the development of better delay management mechanisms. Flight delay prediction has been the topic of several previous efforts. Jetzki [1] studied the propagation of delays in Europe, with the goal of identifying the main delay sources. Tu et al. [2] developed a model for estimating flight departure delay distributions, and used the estimated delay information in a strategic departure delay prediction model. Yao et al. [3] focused exclusively on downstream delays caused by aircraft, cockpit and cabin crew connectivity. By contrast, Bratu and Barnhart [4] focused on the impact of delays on passengers. Recently, Pyrgiotis et al. [5] have considered delay propagation in a network of airports using a queuing model. Other prediction models [6,7,8] have focused on weather-related delays, and the development of a Weather Impacted Traffic Index (WITI). Xu et al. [9] proposed a Bayesian network approach to estimating delay propagation. Using a system-level Bayesian network, the authors were able to capture interactions among airports. Rebollo et al. [10] characterized and predicted air traffic delays. Wong et al. [11] surveyed the model for flight delay propagation. Tu et al. [12] used statistical approach to estimate flight departure delay distributions. By contrast, the goal of this paper is to predict arrival delay using multiple linear regression approach.

3 Data source

The results presented in this paper were obtained using data from the www.umetrip.com. In order to get the flight information, we develop a crawler to crawl data from the website. We had access to a large dataset of 119432 regular commercial passenger flights performed from November 3, 2015 to March 5, 2016. The database provides detailed data for individual flight by phase of flight, airport weather data, arrival and departure time. There are 25 fields in the dataset, and the main fields are shown in Table 1.

Table 1: Fields and description of dataset.

Field	Description
TimeSeries	Date
FlightNo	Flight number
DepAirport	Departure airport
ArrAirport	Arrive airport
DepTime	Planned departure time
ArrTime	Planned arrive time
FlyingTime	Planned flying Time
ActDepTime	Actual departure time
ActArrTime	Actual departure time
ActFlyingTime	Actual flying time
Routing	Air route
Acft	Aircraft type
DistKm	Route distance
windir	Wind direction
windstrength	Wind power
aqi	Air quality

4 Delay prediction based on multiple linear regression

4.1 Principle of multiple linear regression

In statistics, linear regression is an approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables) denoted X. The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression. In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models.

Linear regression has many practical uses. Most applications fall into one of the following two broad categories: (1) if the goal is prediction, or forecasting, or error reduction, linear regression can be used to fit a predictive model to an observed data set of Y and X values. After developing such a model, if an additional value of X is then given without its accompanying value of Y, the fitted model can be used to make a prediction of the value of Y. (2) given a variable Y and a number of variables X_1, ..., X_p that may be related to Y, linear regression analysis can be applied to quantify the strength of the relationship between Y and the X_j, to assess which X_j may have no relationship with Y at all, and to identify which subsets of the X_j contain redundant information about Y.

4.2 Problem definition

The decisions taken in the management of an airport are often based on common sense and influence several variables, such as flight delay. Reducing this delay presents the advantage of decreasing costs and increasing the quality of the service provided to the passengers. It is thus important to find which variables influence flight delay and use them to predict it. In this context, there are many studies. Some of them treat flight delay prediction as a regression problem, predicting the delay by the minute, and others as a classification problem, predicting a time interval where the delay will fall. The problem in this paper considered is to predict flight arrival delay at a given airport. Given information about a flight that will depart from this airport, the main objective of this paper is to predict its arrival delay by the minute. If the delay time fall into [-∞, 30], it indicates that there is no delay in the flight. If the delay time fall into [30, +∞], it indicates that the flight delay. So in this paper, two types of prediction mechanisms are considered: regression, where the continuous output is an estimate of the arrival delay, and classification, where the output is a binary prediction of whether the arrival delay is more or less than the predefined threshold.

4.3 Our approach

It is important to find which variables influence flight delay and use them to predict it. After careful analysis of the data, we found that there is a close relation between arrival delay and departure delay (shown in Fig. 1). So, we can use departure delay to predict arrival delay. We establish a multiple linear regression model taking the form: $Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3$, where the target variable Y is arrival delay, the predictor variables X_1 is departure delay and X_2 is route distance.

The fields used in the analyses in this paper include the planned departure time and actual departure time, the route distance, the departure airport, the aircraft type and the weather of departure airport. At first, we add a field of departure delay subtracting the planned departure time with the actual departure time. In the training phase, we use departure delay and route distance training model in order to learn the three parameters of β_1 , β_2 and β_3 . In the predicting phase, given information about a flight that will depart from one airport, route distance is already known. If we know the departure delay of the flight, we could use the model predict its arrival delay, so as to determine whether the flight delay. Given information about a flight, at first, we look for the similar data according to the departure airport, the aircraft type and the weather of departure airport in the dataset, and then we cluster the similar data using departure delay and set up a threshold. If the amount of data exceeds the threshold, we calculate the average of departure delay, and use it as the departure delay of the given flight. The detailed computational framework is shown in Fig. 2.

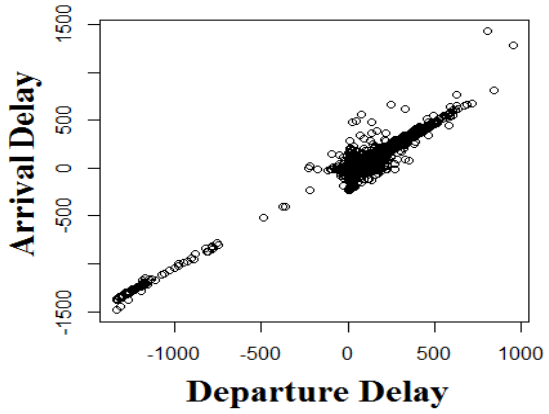

Fig. 1: The relation between arrival delay and departure delay.

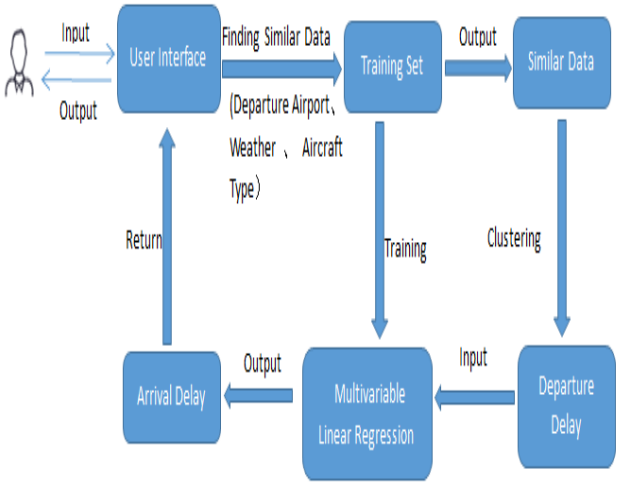

Fig. 2: The detailed computational framework of flight delay.

5 Design and implement of system

In order to predict flight delay, we develop a system. The system includes the crawler and the predictor. The crawler has two fundamental components. The one fetch the webpages and the other one parse the flight information of webpages. We combine depth-first and breadth-first algorithm to get the pages and use regular expression to filter page contents. The crawler is responsible for crawling the flight data from the www.umetrip.com, and the predictor is responsible for training, predicting and testing and the system architecture is shown in Fig. 3. The system was developed by C# language, the development and implemention enviroment is shown in Table 2.

Table 2: The development and implementation environment.

Software Environment	
Development Language	C#
Development Tool	Visual Studio 2012
Operating System	Windows 7
Hardware Environment	
Processor	Inter(R)Core(TM) i5-3470 CPU @3.2GHz
Memory	8.00GB
System Type	64-bit

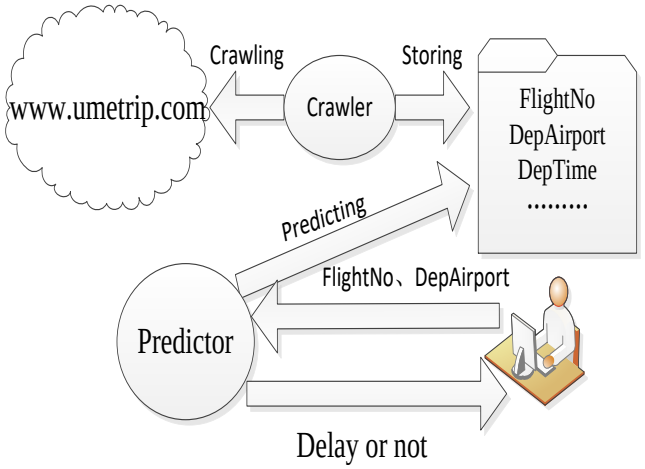

Fig. 3: The basic architecture of system.

6 Experiment

6.1 Data preprocessing

The original dataset contains information for commercial flights in China. Since the data set is extremely large, we extracted a reasonable subset of the data. This reduces the size of our data set to around 100,000 records. For each record, we use secondary data sources to enrich it with information about airplanes and historical weather statistics. At the end of the pre-processing steps, our data set has the following features:

- Time: From November 3, 2015 to March 5, 2016.
- Airport: There are 78 different airports.
- Airline: There are 175 different airlines.
- Flight features: These include the planned times of departure and arrival, the actual times of departure and arrival, the departure and arrival airport, and the distance covered by the flight.
- Airplane features: These include the make and model of the plane.
- Weather features: We obtain historical weather information for the time closest to the departure time. The weather data includes several categorical features indicating the presence of snow, hail, thunder, rain and tornado warnings. It also contains a few numeric features such as wind speed, temperature and humidity.

6.2 Experiment result

In order to predict whether a flight will be delayed or not, we model the problem as a classification with two classes: delayed for flights with delays above 30 minutes, and non-delayed otherwise. We compare our method with Naive-Bayes and C4.5 approach.

We first apply a Naive-Bayes algorithm with ten-fold cross-validation on the entire training set. The Naive-Bayes algorithm runs extremely fast and provides some baseline results, as shown in Tables 3 and 4. The Naive-Bayes results show us that the classifier performance is far better in predicting non-delayed flights than delayed ones. The F-score on predicting on-time flights is 0.75, while that for delays is only 0.57. So, how can we improve the performance of the classifier, particularly in predicting delays?

Then, we train C4.5, which is reputed to a better classifier. Our attempts at using the C4.5 do not greatly improve performance. In fact, the C4.5 performs slightly worse than Naive-Bayes in predicting delays (F-score 0.48). Last, we train our model, experiment result show that our model is better than the Naive-Bayes and C4.5.

Table 3: Performance of classifiers in predicting non-delayed flights.

	Accuracy %	Precision	Recall	F-score
Naive-Bayes	70.2	0.72	0.79	0.75
C4.5	68.3	0.71	0.75	0.65
Our model	79.1	0.79	0.83	0.79

Table 4: Performance of classifiers in predicting delayed flights.

	Accuracy %	Precision	Recall	F-score
Naive-Bayes	70.2	0.72	0.79	0.75
C4.5	68.3	0.71	0.75	0.65
Our model	79.1	0.79	0.83	0.79

7 Conclusion and future work

This paper considered the problem of predicting flight arrival delay and presented a prediction results. The problem was treated as both a regression and an ordinal classification task and a suitable approach, based on the multiple linear regression model, was used to predict the delay. We implemented the model and compare it with Naive-Bayes and C4.5 approach. In future, we will further improve its operational efficiency and accuracy.

Acknowledgements

This work is supported by national education information technology research project under grant No. 156232707, Hubei Provincial Department of Education science research project under grant No. B2015375, Hubei Provincial Vocational Education Institute program under grant No. ZJGA201504. Thanks for anonymous reviewers' valuable comments!

References

[1] Jetzki, M. "The propagation of air transport delays in Europe. Master's thesis", Department of Airport and Air Transportation Research, *Aachen University*, (2009).

[2] Tu, Y., Ball, M.O., Jank, W.S.,. "Estimating flight departure delay distributions – a statistical approach with long-term trend and short-term pattern", *Am.Stat. Assoc. J.*, **103**, pp. 112–125, (2008).

[3] Yao, R., Jiandong, W., Tao, X. "A flight delay prediction model with consideration of cross-flight plan awaiting resources", In: International Conference on Advanced Computer Control, (2010).

[4] Bratu, S., Barnhart, C. "An analysis of passenger delays using flight operations and passenger booking data", *Air Traffic Control Quart*, **13 (1)**, pp. 1–27, (2005).

[5] Pyrgiotis, N., Malone, K.M., Odoni, A. "Modelling delay propagation within an airport network", *Transport. Res. Part C: Emerg. Technol.*, **27**, pp. 60–75, (2013).

[6] Klein, A., Kavoussi, S., Hickman, D., Simenauer, D., Phaneuf, M., MacPhail, T. "Predicting weather impact on air traffic", In: Integrated Communication, Navigation and Surveillance (ICNS) Conference, (2007).

[7] Klein, A., Craun, C., Lee, R.S. "Airport delay prediction using weather-impacted traffic index (WITI) model", In: Digital Avionics Systems Conference (DASC).

[8] Sridhar, B., Chen, N. "Short term national airspace system delay prediction", *Journal of Guidance, Control, and Dynamics*, **32 (2)**, pp. 80-85, (2009).

[9] Xu, N., Laskey, K.B., Donohue, G., Chen, C.H. "Estimation of delay propagation in the national aviation system using bayesian networks", In: 6th USA/Europe Air Traffic Management Research and Development Seminar, (2005).

[10] Rebollo, J. and Balakrishnan, H. "Characterization and prediction of air traffic delays", *Transportation Research Part C: Emerging Technologies*, **44**, pp. 231–241, (2014).

[11] Wong, J.-T. and Tsai, S.-C. "A survival model for flight delay propagation", *Journal of Air Transport Management*, **23**, pp. 5–11, (2012).
[12] Tu, Y., Ball, M. O., and Jank, W. S. "Estimating flight departure delay distributions - A statistical approach with long-term trend and short-term pattern", *Journal of the American Statistical Association*, **103(481)**, 112–125, (2008).