CS5811 Project:
Flight Delay Prediction Based on Neural Networks

Jordon Dornbos
Liang Yan
December 14th, 2014

1 Introduction

Our objective is to develop a neural network model for airplane delay prediction.

This paper describes a neural network model for predicting flight delay

We are motivated to solve this problem because flight delays are a fairly common thing and there seems to be no good source for predicting this. Flight delays are generally only announced about an hour or so before the flight is set to takeoff and they can change unreliably. Our solution can't guarantee 100% accuracy for predicting a delay, but it will attempt to better prepare passengers for any possible delays.

1.1 Problem statement

I believe that almost everyone has had a bad experience with a flight being delayed, especially being delayed for something important. According to the U.S. Department of Transportation, over 20 percent of all flights arrive late.[4] It is a major problem for the airplane company and also for us, passengers.

Most of delays are caused by three reasons. First, and most important, is weather. Half of delays are actually caused by weather[2]. Second, it could be some mechanical problem[1], which is mostly decided by the plane features such as the make, model, age and so on. Third, it could be caused by a scheduling problem. If the last flight is delayed, this flight could be delayed also. [3]

1.2 State of the Art

Dr. Rebollo and Balakrishnan[3] proposed amodel uses Random Forest (RF) algorithms, based on temporal and spatial delay, also the local arrival or departure delay situation, they proposed this new network delay prediction model. The predictive performance of the model is evaluated using the 100 most delayed OD pairs in the NAS: the results show that given a 2-hour prediction horizon, the average test error across these 100 OD(origin-destination) pairs is 19% when classifying delays as above or below 60 min.

Dr. Lu and his team predict flight delays by creating a decision tree from a large database. [4]This solution is good at making a simple tree to follow in order to make predictions, but a new tree

needs to be created once more data is collected. Using neural networks, however, the graph stays the same, the weights are just learned over time. Our solution will be learning how each input parameter affects the delay probability. This should therefor be more accurate than creating a graph of the most likely outcome based on a series of tests, such as the decision tree provides.

2 Data collection

2.1 Airline Delay Root Cause Analysis

According to the statics of BTS on flight delays since June 2003, there are five main reasons for a flight delay:

First, air carrier: the airplane was delayed or canceled because of a problem within company, such as aircraft maintenance, cleaning, fueling or crew problems.

Second, extreme weather: the plane was canceled because of significant conditions like a tornado, blizzard or hurricane.

Third, National Aviation System: because of some scheduling problem, it could cause the flight to get delayed or even canceled.

Last, security reasons: delays or cancellations caused by evacuation of a terminal or concourse, or re-boarding of aircraft because of a security breach.

2.2 Data collecting

We are using Python as the main programming language and we get data from the U.S. Department of Transportations Bureau of Transportation Statistics. It is a very big data set, and is formatted as the following:

http://stat-computing.org/dataexpo/2009/the-data.html

Name	Description
1	Year 1987-2008
2	Month 1-12
3	DayofMonth 1-31
4	DayOfWeek 1 (Monday) - 7 (Sunday)
5	DepTime actual departure time (local, hhmm)
6	CRSDepTime scheduled departure time (local, hhmm)
7	ArrTime actual arrival time (local, hhmm)
8	CRSArrTime scheduled arrival time (local, hhmm)

```
9       UniqueCarrier   unique carrier code
10      FlightNum       flight number
11      TailNum plane tail number
12      ActualElapsedTime       in minutes
13      CRSElapsedTime  in minutes
14      AirTime in minutes
15      ArrDelay        arrival delay, in minutes
16      DepDelay        departure delay, in minutes
17      Origin  origin IATA airport code
18      Dest    destination IATA airport code
19      Distance        in miles
20      TaxiIn  taxi in time, in minutes
21      TaxiOut taxi out time in minutes
22      Cancelled       was the flight cancelled?
23      CancellationCode        reason for cancellation (A = carrier,
B = weather, C = NAS, D = security)
24      Diverted        1 = yes, 0 = no
25      CarrierDelay    in minutes
26      WeatherDelay    in minutes
27      NASDelay        in minutes
28      SecurityDelay   in minutes
29      LateAircraftDelay       in minutes
```

We will construct a MySQL database which will use the flight time and flight number as the main keys. We also record its origin and destination airport, and a variable if it is delayed (cancellation is a delay too), and the tail number of the plane.

We also get the weather information from from National climate data center (NCDC). The weather details were available only for 180 airports of the 315 airports for which the flight stats were available.

We use the airplane departure time to find the relative weather information, which we'll save in the database.

2.3 Data preprocessing

According to the reasons above, we focused on the following problems:

1. When predicting if an airplane will be delay, we check the time first, according to the statics. Airplane delay has a strong connection with the departure time

2. Airports: different airports have different air traffic volume and different schedule constraints which could be reflected in the NAS reason

3. Carrier: different companies have different delay rate, we will calculate its delay rate from last year as input.

4. The airplane itself: we will use the tail number as the unique id, and calculate its delay rate from the last year as its input.

5. weather: we will change the weather data to a percentage according to how much of a delay it will cause.

Since the effectiveness of security is very limited, we just ignored it in this model.

3 Methodology

We used past flight information to train and test our neural network and experimented with changing parameters of the neural network to figure out what values work best. We used an automated testing system that varies the amount of hidden layers, and nodes per hidden layer. We then used this output to find the best set of parameters for the neural network.

3.1 Artificial Neural Network Model

The input to the neural network is normalized data that we get from our preprocessed data sources. The neural network processes this input using a sigmoid activation function that gives an output from 0-1. There are a set number of hidden layers, nodes per hidden layer, and randomized importance weights between nodes. The stopping criteria for learning is a set number of iterations, which can be adjusted. For our testing we set this value to 10,000 iterations. Once the known values are fed forward, back propagation is used to adjust the weights based on the output error. We also used simulated annealing to change the learning rate, so that it makes smaller changes to the weights as time goes on. We then apply a threshold of 0.5 to the output layer, to get our prediction. Values above 0.5 will result in a delay prediction, while values less than or equal to 0.5 result in an on time prediction.

We trained the neural network by feeding it data on past flight data, and back-propagating the correct value once the values have been fed through the network. Over time the neural network will learn the importance of each input, which allows it to make accurate predictions. The datasets that we've used are from the U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics (BTS). We used the data from 2004 to train the network, and the data from 2007 to test the network for accuracy.

3.2 Experiment Setup

For our experiment we wrote the program in Python and imported the data from CSV files. The CSV files used are a combination of raw data from the Bureau of Transportation Statistics, as well as normalized data that we have created. This normalized data is derived from the Bureau of Transportation Statistics' data.

Since the dataset is very large, consisting of millions of flight records, we use a random subset of this dataset. We add every 1000th data point, which results in over 7,000 flight records. Adding every 1000th data point avoids getting biased data based on how the data is sorted in the CSV file.

From this subset we shuffle it to avoid any biases from the sorted CSV file. We also order it so that every other values is either a delayed flight or an on-time flight. This is done to avoid getting a biased result from the learning phase. If the end of the dataset is heavily weighted towards one value, it can skew the weights towards that value. This can cause the network to make predictions that favor this value after the learning phase has completed.

4 Result

As mentioned previously, we trained the neural network using flight data from 2004, and then tested the network using flight data from 2007. From our testing we found that the network works best with 3 hidden layer and 10 nodes per hidden layer. We found that this correctly predicted 1,409 delayed flights and 349 on-time flights. It incorrectly predicted 116 delayed flights and 1,176 on-time flights. This results in an average accuracy of 57.6%. This is similar to other methods that we have investigated. Our threshold for delayed or on time was 50%. Changing this could improve both the accuracy and the high amount of incorrect on time flight predictions.

5 Conclusion

References

[1] Sina Khanmohammadi, Chun-An Chou, Harold W. Lewis, and Doug Elias. A systems approach for scheduling aircraft landings in jfk airport. In *Fuzzy Systems (FUZZ-IEEE), 2014 IEEE International Conference on*, pages 1578–1585, July 2014.

[2] A Klein, C. Craun, and R.S. Lee. Airport delay prediction using weather-impacted traffic index (witi) model. In *Digital Avionics Systems Conference (DASC), 2010 IEEE/AIAA 29th*, pages 2.B.1–1–2.B.1–13, Oct 2010.

[3] Juan Jose Rebollo and Hamsa Balakrishnan. A network-based model for predicting air traffic delays. In *5th International Conference on Research in Air Transportation (ICRAT 2012)*, May 2012.

[4] Lu Zonglei, Wang Jiandong, and Zheng Guansheng. A new method to alarm large scale of flights delay based on machine learning. In *Knowledge Acquisition and Modeling, 2008. KAM '08. International Symposium on*, pages 589–592, Dec 2008.