A REMIXED BAYESIAN NETWORK BASED ALGORITHM
FOR FLIGHT DELAY ESTIMATING

Yujie Liu¹, Han Wu²

¹College of Computer Science and Technology
Civil Aviation University of China
Tianjin, P.R. CHINA

²Department of Mathematics and Statistics
Minnesota State University
Mankato, MN, USA

Abstract: A new Bayesian Network algorithm is proposed in this paper. When seeking more accurate results, this new algorithm, Negotiating Method with Competition and Redundancy (NMCR), has bigger scale in structure than other network models we proposed. Time needed for network training and speed in testing show that NMCR works well in estimating of arrival flight delay, especially in flight chains mainly operated in China.

AMS Subject Classification: 62F15, 68U20, 81T80
Key Words: Bayesian network, flight delay, modeling and estimating, remixed algorithm

1. Introduction

Flight delay is a global problem that concerns industrial experts for years, and much research has been done through different approaches[1] [2] [3]. Bayesian
Network (BN), a machine learning method based on graph theory and probability theory, is an effective tool used to model and estimate unstable complex problems with plenty records [4]. Flights’ delay is exactly this kind of problem. In recent years, we have proposed a few BN models with different improved algorithms, which are based on parameter learning, structure learning, and some mixed algorithms, respectively (see [5]-[8],[10] and [11]). There are still many aspects needed to be improved in these algorithms and models.

<table>
<thead>
<tr>
<th>flight nd</th>
<th>ac_type</th>
<th>dep apt</th>
<th>arr apt</th>
<th>ac_id</th>
<th>std</th>
<th>seats</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA1827</td>
<td>PEK</td>
<td>WEH</td>
<td>07:45:00</td>
<td>167</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CA1828</td>
<td>WEH</td>
<td>PEK</td>
<td>10:05:00</td>
<td>167</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CA1546</td>
<td>PEK</td>
<td>YNT</td>
<td>12:20:00</td>
<td>167</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CA1143</td>
<td>PEK</td>
<td>DSN</td>
<td>18:20:00</td>
<td>167</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CA1144</td>
<td>DSN</td>
<td>PEK</td>
<td>20:50:00</td>
<td>167</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(a) Two Flight Chains Selected From a Flight Schedule

(b) The Delay Propagating in a Flight Chain

Figure 1: Flight Chain

2. Flight Delay in Flight Chain

A Flight Chain (FC) refers to a chain structure that is a part of an airline’s flight schedule. The Flight Chain contains flights in chronological order. A sample
Flight Chain is shown in Figure 1(a). Most of the time these flights are executed by the same aircraft that can be distinguished by “ac_id” easily and they are usually operated by the same airline company. We call this relationship between flights In-Flight-Chain-Aircraft-Correlation, one of the most three correlations between flights [8]. The sketch map of a Flight Chain is shown in Figure 1(b).

In this paper, we are to find a suitable Bayesian Network architecture and algorithm to pursue more accurate results in modeling and estimating the flight delay in a Flight Chain (FC).

### 2.1. Flight Delay in Flight Chain

Theoretically the flight delay may happen in any airport for various reasons. The airports all over the world are different. Some are busy while some are not. The flight delay, especially the initial delay, tends to happen in busy hub-airports, the place we used to call it the hot or congested nodes in a FC. The initial delay may be arrival delay, or departure delay. When the initial flight delay happens, it often propagates through FC, which leads to so called delay propagation. Therefore more delay will happen and more flights will be delayed [6] [7]. Flight delays may be controlled or even reduced, only if we can estimate accurately the initial delay time in the first place. Therefore we focus more on the arrival delay records in some busy hub airports in a Flight Chain.

### 2.2. Previous Work on Algorithms for Delay Estimating

An improved algorithm named Target-Fixed Stochastic-Ordered K2 (TSK2) has been put forward to establish a Bayesian Network in our previous work [10]. It is an improved K2 algorithm. K2 is an algorithm that heuristically searches for the most probable belief network structure given a database [9]. The Bayesian network built on K2 for flight delay estimation is shown in Figure 2 (a). The detail can be found in [10].

TSK2 has been proven to be effective in modeling and estimating of the flight delay. The models trained by TSK2 provide a successful topologic for estimating the flight delay. Tests on different data verify that TSK2 is better than classic K2 both in speed and accuracy. The Bayesian network built on TSK2 for flight delay estimation is shown in Figure 2(b). The detail can be found in [11].

The new algorithms including TSK2 are proposed to decrease the error rate as much as possible. However, there are always some errors existing. We have
analyzed the results and focused on the error part. We have checked these testing data, and have found there are always some special flights that could not be estimated correctly by BN models even if in different algorithms. Some samples are listed in Figure 2(c).

A similarity of these flights record is that most of them are not regular passenger aircrafts. Some are chartered aircrafts while some are cargo aircrafts, and so on. Then we try to find a way to build a model that can accommodate these unusual flights, other than just treat them as useless outliers.
3. Negotiating Method with Competition and Redundancy (NMCR)

Following the idea stated in previous section, a new algorithm, Negotiating Method with Competition and Redundancy (NMCR), is proposed. Based on NMCR, a new Bayesian Network, NMCR net, is trained, tested, and then established.

NMCR net includes two subnets and one negotiating node. The algorithms of two subnets are all based on TSK2. The two subnets, N1 and N2, are shown in Figure 3(a). The training data set for N2 is different from the one for N1. It does not include those data that can be accurately estimated by N1. N2 tends to have totally different topologic from N1.
3.1. Training of NMCR

H is a control node in charge of negotiating between two subnets. It determines whether two subnets will be redundant or competitive. It plays an important role in NMCR network. To increase the operating speed of the network, classic statistic method is chosen in node H.

The work needed to be done in training includes the following three parts:

1. To build N1 network model based on TSK2. The data used: all the training data.

2. To build N2 network model based on TSK2. The data used: the data that have been incorrectly estimated in N1.

3. To train negotiating node H. The data used: all the results from N1 with successful or fault marks.

The main feature of NMCR is that many redundancy nodes are created during the training period.

3.2. Testing of NMCR

In testing period, these redundancy nodes have two states, competition or redundancies, which means whether the redundancy nodes work or not. And the behavior of these redundancy nodes is under the control of H. After training, node H can do the negotiating work when a new data go through it. You may see this process clearly in Figure 3(b). For example, when a new record D enters this model through node H, two probabilities $P_1$ and $P_2$ may be calculated for D. $P_1$ is the probability that D can be estimated successfully in N1 while $P_2$ is such probability in N2. Compare $P_1$ and $P_2$ as follows,

\[
\begin{cases} 
|P_2 - P_1| > 0.1 & \text{model works in redundancy} \\
|P_2 - P_1| < 0.1 & \text{model works in competition}
\end{cases}
\]

When $|P_2 - P_1| > 0.1$, NMCR model works in redundancy. If $P_2$ is bigger, N2 will work, and N1 is redundant. On the other hand, if $P_1$ is bigger, N1 will work and N2 is redundant. When $|P_2 - P_1| < 0.1$, N1 and N2 work together. By comparing the results from two subnets, the output with the biggest probability will be chosen as the final estimating result.
3.3. The Algorithm and Structure of NMCR

We apply this new algorithm to model arrival delays for one busy hub airport, and we get the network shown in Figure 4.

After training, we use another data set for testing. We obtain the estimating result from NMCR, shown in the upper panel of Figure 5.

We see that the test data set has more than 16,000 records. This NMCR result shows that the accuracy rate of estimating is 90.512%, and the error rate is 9.488%.

3.4. The Main Difficulty in NMCR

The main difficulty of NMCR net is the relatively slow speed in training and testing. From the analysis of NMCR, we know that the scale of NMCR net is much bigger than other Bayesian Networks if trained by the same data set. It
shows clearly in Figure 4 and Figure 2 (a) and (b) that NMCR is complicated. For other Bayesian Networks, such as K2 or TSK2, the node number usually equals to the attributions of the data. But the nodes of NMCR net are more than twice as much as the attributions number.

In a regular Bayesian Network, as the network scale grows, the calculation shows a geometrical scale growth, since every node needs to count every status of all the nodes above itself. Thanks to the function of competition and redundancy, nodes in the NMCR net do not always work together in the testing period. In the training period, N2 is built after N1. Therefore the time scale increases linearly. In the lower panel of Figure 5, we can see that the modeling speed of NMCR net is fast, which is about 0.34 second. The time changes slightly for different data sets. This example shows that NMCR algorithm may solve the problem of relatively slow speed.
In Figure 5, the “error” column is to mark the estimation errors. It is left blank if the estimating result is correct. Five columns of “Probability distribution” includes all the probabilities in estimating. There are five degrees of flight delay in training and testing data. “*” indicates the max probability in estimating.

4. Conclusions

To accommodate some unusual fights in estimating of BN, we propose a new algorithm, named NMCR, to model and estimate the arrival flight delay for a busy hub airport. NMCR is a remixed algorithm, and the Bayesian Network built upon it has complicated topologic and more than twice nodes as many as other BNs. Due to the competition and redundancy, the speed of NMCR is proven acceptable. And owing to the second subnet N2, NMCR can improve the correct rate for any source data sets based on TSK2.

Acknowledgments

This work is supported by the National Nature Science Foundation of China under Project 61179046 and 60879015.

References


Appendix

Some technical definitions:

a) Arrival Delay: Arrival Delay refers to the delay time of flight arrival.

b) Departure Delay: Departure Delay means the delay time of flight departure.
c) Delay Propagation: Delay Propagation is a special character of flight delay. Delay transferring from one flight to another, we call this phenomenon "Delay Propagation".

d) Initial Delay: the first delay happens in a Flight Chain, it is the original source of delay propagation.

e) Aircraft-correlation: There are three main correlations among flights: aircraft-correlation, pilot-correlation, and crew-correlation. These correlations are considered when the airlines make flight schedule in every new season.