FLIGHT DELAY STATE-SPACE MODEL BASED ON GENETIC EM ALGORITHM

Chen Haiyan[1], *Wang Jiandong*[1], *Xu Tao*[2]

(1. College of Computer Science and Technology, NUAA, 29 Yudao Street, Nanjing, 210016, P. R. China;
2. College of Computer Science and Technology, Civil Aviation University of China, Tianjin, 300300, P. R. China)

Abstract: Flight delay prediction remains an important research topic due to dynamic nature in flight operation and numerous delay factors. Dynamic data-driven application system in the control area can provide a solution to this problem. However, in order to apply the approach, a state-space flight delay model needs to be established to represent the relationship among system states, as well as the relationship between system states and input/output variables. Based on the analysis of delay event sequence in a single flight, a state-space mixture model is established and input variables in the model are studied. Case study is also carried out on historical flight delay data. In addition, the genetic expectation-maximization(EM) algorithm is used to obtain the global optimal estimates of parameters in the mixture model, and results fit the historical data. At last, the model is validated in Kolmogorov-Smirnov tests. Results show that the model has reasonable goodness of fitting the data, and the search performance of traditional EM algorithm can be improved by using the genetic algorithm.

Key words: flight; delay; predictions; dynamic data-driven application system; genetic EM algorithm

CLC number: TP391; U8 **Document code**: A **Article ID**: 1005-1120(2011)03-0276-06

INTRODUCTION

As a result of excessive demand for air transportation, the flight delay becomes an urgent problem that exacerbates national transportation bandwidth limitations. Over the past decade, researches were focused on analyzing flight delay factors, predicting delay and propagation, and mitigating delays and other issues[1-4]. Deterministic models are commonly used in delay prediction. For example, one of the models is to estimate delays according to flight schedule. Models like this usually ignore random factors such as unexpected events and queuing. Prediction models based on random density functions of seasonal trends, daily propagation and daily delay[5-6], that to a certain extend reflect the overall patterns of flight delays, are insufficient in capturing variations in individual flight delay.

Real-time prediction of flight delay is essential in the state estimation process for a dynamic system. Flight operation process is monitored in order to collect data in real time, which provides an opportunity to apply dynamic data-driven application system (DDDAS)[7] that can dynamically employ prediction to control and guide the measurements, and in reverse, can dynamically steer the prediction based on the measurements. DDDAS promises more accurate analysis and prediction, more precise controls, and more reliable outcomes, which can improve advance prediction capabilities of prediction systems. The challenge in the problem remains in establishment of the delay state-space model, which is the foundation in applying the dynamic data-driven approach. P. Wang[8] presented a simple recursive model based on delay propagation. In the model, P. Wang demonstrated a linear relationship a-

Foundation item: Supported by the High Technology Research and Development Programme of China (2006AA12A106).

Received date: 2010-09-07; **revision received date**: 2010-12-27

E-mail: chenhaiyan@nuaa.edu.cn

mong system states while ignored the effective pattern of uncertainties. In this paper, a recursive model is further improved with the use of an explicit expression to calculate flight delay caused by random factors. Delay information is feedback to the state-space model as the system input. In order to search for maximum likelihood estimates of parameters in the model, the genetic algorithm (GA) is combined with the traditional expectation-maximization(EM) algorithm to avoid the local maximum problem. Performance comparison between the model and the genetic EM algorithm is given as well.

1 STATE-SPACE MODEL OF FLIGHT DELAY

1.1 Delay propagation of flight

From departure at an airport to arrival at the destination, an aircraft accomplishes a flight task. For efficiency and cost considerations, an aircraft should perform multiple tasks consecutively each day. Assume d denotes a departure event and a an arrival event. Then the discrete event sequence of an aircraft performs in a day can be written as $d_1a_1d_2a_2\cdots d_na_n$, where the state of the next event only depends on the state of the current event, and not on the state of the past event. The discrete events sequence is a Markov chain. Therefore, the relationship among states can be represented in a state-space model.

1.2 State-space model of flight delay

The state-space model of flight delay based on the recursive model[8] can be expressed as

System model

$$x_{i+1} = x_i + u_i + w_i \tag{1}$$

Measurement model

$$y_i = x_i + v_i \tag{2}$$

where x_i denotes the state variable, u_i the system input, y_i the measurement, w_i and v_i denote the process and measurement noise, respectively, and both are random white noises. The system model (1) describes the evolution of the state variables over the sequence, whereas the measurement model (2) represents how measurements relate to the state variables. If an aircraft accomplish n flight tasks, then we have $i=1,\cdots,2n$. When i is an odd number, x_i denotes a departure delay state or an arrival delay state, vise versa.

Since the flight delay in this paper represents the difference between the actual flight time and the scheduled flight time. Random factors such as weather, baggage check-ins, and mechanical failures may result in a delayed flight. On the other hand, an early flight task completion is achievable through planning methods and strategies. Flight delays caused by these uncertainties can be added to the model as u_i. Additionally, air turnaround time and ground turnaround time correspond to two uncorrelated processes. Values of u_i for different models should be estimated in two delay states. However, the relationship between the uncertainties and the flight delays is not represented by any mathematical models, which leaves the calculation of u_i as a key problem in establishment of the state-space model.

1.3 Modeling of system input

In general, x_i is the departure delay from an upstream airport, u_i is represented as the delay in air. When $u_i<0$, it is actually denoted as flight time compensation. Earlier statistics show that the longer itinerary duration a flight is to take, the more compensation the flight can obtain. And the longer itinerary duration impacts on the final status of the arrival delay. As a result, a more effective way to represent u_i is given as follows

$$u_i = fs_i * r_i \tag{3}$$

where fs_i denotes the scheduled flight time between airports, r_i the delay of per scheduled flight time, or delay rate. Table 1 shows delay rates in percentage at different levels extracted from the historical flight data.

Nearly 85% of flights obtain compensation in some levels, while 15% of flights end in flight delays. The delay rates vary significantly in distribution, decreasing sharply as a function of the distance from the center. The statistic result suggests us to use a finite mixture model to describe

Table 1 Percentages of delay rates at different levels

Rate	Percentage/%
(−1,−0.4)	0.2
[−0.4,−0.2)	10.7
[−9.2,0]	73.8
(0,0.2]	14.4
(0.2,1)	0.9

the delay rate distribution. Finite mixture distribution model[9] is a mathematical method to model the generic random phenomena. Long-term empirical results show the high adaptability of this method. The density distribution g of delay rate is modeled as a function with m mixed components. The mixture density of the ith point is written as

$$g(r_i|\Theta) = \sum_{j=1}^{m} \alpha_j \psi_j(r_i|\theta_j) \quad (4)$$

where $\Theta = (\alpha_i, \cdots, \alpha_m, \theta_1, \cdots, \theta_m)$ denotes the parameter vector, $\alpha_j (\alpha_j \in [0,1], \sum_{j=1}^{m} \alpha_j = 1)$ the mixing weight of the jth component, and $\psi_j(r_i|\theta_j)$ the density function of the jth component depending on parameter θ_j. In this paper, we assume that g is a normal mixture model. And θ_j is denoted as $\theta_j = (\mu_j, \Sigma_j)$, where μ denotes the mean and Σ the covariance matrix.

In the finite mixture model of data set $\boldsymbol{r} = (r_1, r_2, \cdots, r_n)$, r_i is assigned to the most possible component. Then, a label vector set of r_i, $\boldsymbol{z} = (\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_n)$ is obtained. If r_i belongs to the kth component, then $z_{ik}=1$ and the rest label variants are set to 0. Parameter vector Θ is estimated to obtain $\boldsymbol{z}$. And the log-likelihood of Θ is given as follows

$$\log L(\Theta|\boldsymbol{r}) = \sum_{i=1}^{n} \sum_{j=1}^{m} \{\log \alpha_j + \log \psi_j(r_i|\theta_j)\} \quad (5)$$

2 PARAMETER ESTIMATION BASED ON GENETIC EM ALGORITHM

The EM algorithm[10] is the most popular and effective method for parameter estimation. The traditional EM algorithm is an iterative two-step procedure: E-step and M-step. The E-step calculates the expectation of the log-likelihood on the observed data $\boldsymbol{r}$ and the current value of Θ. The M-step updates the corresponding estimate of Θ. After a certain number of iterations, the algorithm obtains the local optimal value of Θ. In order to avoid the local maximum problem associated with the traditional EM algorithm, calculation mechanism of GA can be applied to EM to find the global optimum. The combination of GA and EM is known as genetic EM algorithm[11].

In this paper, the fitness function used in the genetic EM algorithm is the log-likelihood function defined in Eq. (5) and calculation stops when improvement of the fitness function value decreases below a given threshold. The procedure of the genetic EM algorithm is shown as follows

```
Initial: oldChrom←(Θ_1^0,⋯,Θ_k^0)
        EMrate←1E-8;
        bestFit←10000;
        oldFit←100;
    while(bestFit-oldFit)> EMRate
      fitV←Evaluation (oldChrom, r);
       newChrom ← Selection (oldChrom, fitV, p_s);
       newChrom ← Crossover (newChrom, k, p_c);
      newChrom←Mutation(newChrom,p_m);
      newChrom←EM(newChrom, r);
      oldFit←bestFit;
      bestFit←max(fitV);
      newChrom←sortByMiu(newChrom);
      oldChrom←newChrom;
    end
```

3 CASE STUDY AND VALIDATION

The flight operation data used in this case study is provided by a domestic airline. Information like arrival delay, upstream delay propagation and delay rate is extracted from the experimental data which is also divided into several groups categorized by operating date, testing set (only one set), and training set (excepting the testing set). Parameters are estimated using the genetic EM algorithm on the training set. The fitness of the model is validated on the testing set.

3.1 Density estimation of delay rate

Density estimation of delay rate is implemented in Matlab7.1. The density distribution of the original delay rate is shown in Fig. 1, where the distribution represents a mixture of normal distributions rather than a single normal distribution. Assuming component number $m=1,2,3,4$, we obtain one single model and three mixture models after parameter estimation. As a result, Fig. 2 shows a fitted distribution with two components.

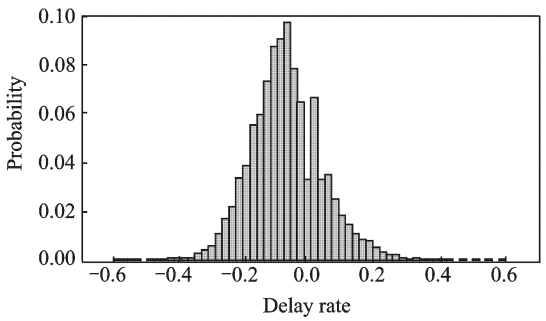

Fig. 1 Density distribution of original delay rate

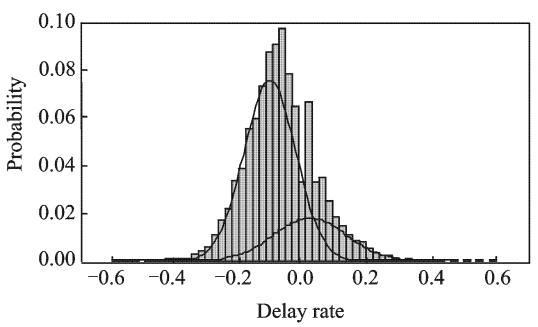

Fig. 2 Fitted distribution with two components

3.2 Fitness test of model

Since the normal mixture models are mixtures of normal distributions, general test methods cannot be directly applied to fitness test for the model. Therefore, a hypothesis test based on Kolmogorov-Smirnov method is used in the test with steps shown as follows.

(1) Generate a number of random samples according to the density function of the mixture model, where the sample set is denoted as $\boldsymbol{X}_1$ and the testing set is denoted as $\boldsymbol{X}_2$.

(2) Give a null hypothesis H_0: in which $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are drawn from the same continuous distribution.

(3) Run the Matlab function″ (h, p) = ktest2($\boldsymbol{X}_1$,$\boldsymbol{X}_2$)″to find whether the distributions are the same at the 5% significance level. If the significance level equals or exceeds the p-value then we have $h=1$, otherwise $h=0$. Reject H_0 if $h=1$ or accept the null hypothesis if $h=0$.

The Results from all four tests on these models are listed in Table 2. The null hypothesis is accepted when $m=2$. Therefore, for the case study, the normal mixture model with two components has the best fitness.

Table 2 Results of model tests

m	1	2	3	4
h	1	0	1	1
p	0.006	0.269	0.008	0.000 1

3.3 Performance validation of genetic EM algorithm

The performance of model is validated in the calculation through the comparison between the genetic and the traditional EM algorithms. On the same stop criteria, the log-likelihood values produced in all iterations from the two EM algorithms with $m=3$ are collected and shown in Fig. 3. In each step, the genetic EM algorithm achieves the better log-likelihood value, which represents the higher effectiveness.

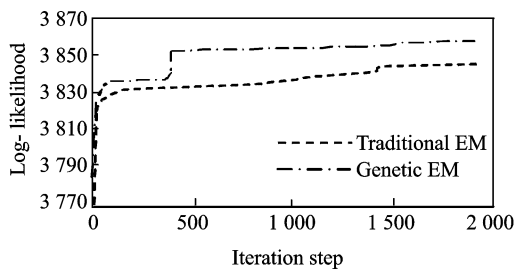

Fig. 3 Log-likelihood values of genetic EM and traditional EM

Additionally, the total iteration numbers of the two EM algorithms, denoted as m, are compared in Table 3. Results show that the iteration number of traditional EM algorithm increases significantly with larger m. The iteration number increases slightly in the genetic EM algorithm.

When the algorithm preparation time is ignored, the genetic EM algorithm can achieve the faster convergence and maintain the higher accuracy than the traditional EM algorithm.

Table 3 Iteration Steps with increasing m

Algorithm	m		
	2	3	4
Traditional EM	968	1 865	3 053
Genetic EM	936	1 734	972

4 CONCLUSION

In this paper, a flight delay state-space model is proposed based on the delay propagation. In the model, delay from the upstream event is denoted as a current state, while the delay caused by other uncertainties is denoted as the system input. System inputs are produced using different models when two delay states are estimated. The modeling process is demonstrated in detail. The genetic EM algorithm is used to find the global optimal estimates of the parameters in the normal mixture model of random delay. Case study and model validation are carried out on real flight data. Results show that the model has an excellent fit to the real data in both the mixture density distribution calculation and the Kolmogorov-Smirnov test. In conclusion, the traditional EM algorithm can be optimized and become more efficient by using GA method in finding the global optimum. Most importantly, the flight delay state-space model proposed in this paper can make it possible to apply DDDAS to the air transportation industry in the near future. DDDAS architecture for flight delay prediction can be established based on this computational model, together with the advanced measurement infrastructure and information technology.

ACKNOWLEDGEMENT

Authors would highly appreciate the anonymous domestic airline, which provided historical flight information.

References:

[1] Abdelghany K F, Shah S S, Raina S, et al. A model for projecting flight delays during irregular operation conditions [J]. Journal of Air Transport Management, 2004, 10(6): 385-394.

[2] Hsu C L, Hsu C C, Li H C. Flight delay propagation, allowing for behavioral response [J]. International Journal of Critical Infrastructures, 2007, 3(3/4): 301-326.

[3] Ding Jianli, Yu Yuecheng, Wang Jiandong. A model for predicting flight delay and delay propagation based on parallel cellular automata [C]//ISECS International Colloquium on Computing, Communication, Control and Management. Washington D C, USA: IEEE, 2009: 70-73.

[4] AhmadBeygi S, Cohn A, Lapp M. Decreasing airline delay propagation by re-allocating scheduled slack [J]. IIE Transactions, 2010, 42(7): 478-489.

[5] Tu Y F, Ball M, Jank W. Estimating flight departure delay distributions—A statistical approach with long-term trend and short-term pattern [J]. Journal of the American Statistical Association, 2008, 103 (481): 112-125.

[6] Abdel-Aty M, Lee C, Bai Y Q, et al. Detecting periodic patterns of arrival delay [J]. Journal of Air Transport Management, 2007, 13(6): 355-361.

[7] Darema F. Introduction to the ICCS 2007 workshop on dynamic data driven applications systems [C]//International Conference on Computational Science. Berlin, Heidelberg: Springer-Verlag Press, 2007: 955-962.

[8] Wang P T R, Schaefer L A, Wojcik L A. Flight connections and their impacts on delay propagation [C]//Digital Avionics Systems Conference. Washington D C, USA: IEEE, 2003, 1(5. B. 4): 1-9.

[9] McLachlan G, Peel D. Finite mixture models [M]. New York: John Wiley, 2000.

[10] Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm [J]. Journal of the Royal Statistical Society: Series B (Methodological), 1977, 39(1): 1-38.

[11] Pernkopf F, Bouchaffra D. Genetic-based EM algorithm for learning gaussian mixture models [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1344-1348.

基于遗传EM算法的航班延误状态空间模型

陈海燕[1] 王建东[1] 徐 涛[2]

(1. 南京航空航天大学计算机科学与技术学院,南京,210016,中国;
2. 中国民航大学计算机科学与技术学院,天津,300300,中国)

摘要:航班运行过程的高度动态性和随机性,航班延误因素的复杂性和不确定性导致航班延误实时预测成为难题。控制领域的动态数据驱动方法为该问题提供了一种解决方案。然而,要想运用动态数据驱动方法,首先必须建立航班延误状态空间模型来表示系统状态之间、状态与系统输入输出之间的关系。本文对单机延误事件序列进行了分析,创建了一种航班延误状态空间模型,并对其中的输入控制量进行了重点建模。在历史航班运行数据集上,采用遗传EM算法对模型参数进行了极大似然估计,并同时验证了遗传EM算法在优化参数估计和提高计算效率方面的优势。最后,采用Kolmogorov-Smirnov方法对模型实例进行了假设检验,检验结果表明,所选模型具有较好的拟合优度。

关键词:航班;延误;预测;动态数据驱动应用系统;遗传EM算法

中图分类号:TP391;U8

(Executive editor: Zhang Huangqun)

基金项目:国家高技术研究发展计划("八六三"计划)(2006AA12A106)资助项目。