An Adaptive Online Learning Model for Flight Data Cluster Analysis

Weizun Zhao
Department of System Engineering and Engineering Management
City University of Hong Kong
Hong Kong, People's Republic of China
wzzhao6-c@my.cityu.edu.hk

Fang He
School of Aeronautics and astronautics
Shanghai Jiao Tong University
Shanghai, People's Republic of China
fanghe@cityu.edu.hk

Lishuai Li
Department of System Engineering and Engineering Management
City University of Hong Kong
Hong Kong, People's Republic of China
lishuai.li@cityu.edu.hk

Gang Xiao
School of Aeronautics and astronautics
Shanghai Jiao Tong University
Shanghai, People's Republic of China
xiaogang@sjtu.edu.cn

Abstract— **Safety is a top priority for civil aviation. To help airlines further improve safety, various clustering-based methods were developed to better understand their current flight operations and detect unknown risks from onboard flight data. However, existing methods can only be carried on historical data in batches, resulting in its inability to update and adjust as new data come in. New onboard flight data related to anomaly detection are generated at airlines every day. The addition of new data will inevitably cause changes in the clustering results. Yet it would be computational costly to run clustering on all data as they accumulate. Therefore, anomaly detection methods that allow real-time update of cluster models as new data come in are more practical for airlines. This paper presents a reinforcement learning method to identify common patterns in flight data via cluster analysis and update its clusters as new data come in. This method is based on Gaussian Mixture Model (GMM) and uses online (recursive) expectation-maximization (EM) algorithm to update clustering results over time. An initial result of clusters can be obtained by performing GMM-based clustering on historical flight data. Then, as new data come in, the parameters of GMM are updated via an online EM algorithm. By recording the GMM parameters, the method can also track changes in clusters over time. We demonstrated the proposed method using Flight Data Recorder (FDR) data from real operations of an airline. The evolution of clusters was observed as new batches of flight data are fed into the proposed method.**

Keywords—Gaussian mixture models, adaptive online clustering, expectation maximization, flight data

I. Introduction

In the past, the improvement of aviation safety was mostly based on the analysis of accidents and lacked of the ability to effectively detect and deal with the unknown safety hazard. In recent years, the aviation industry has adopted many data-based risk identification methods. The digital flight data recorder (FDR) data is used by many airlines for routine analysis to identify risks. In order to better identify factors in FDR data that cannot be artificially identified, some methods based on cluster analysis have been developed, such as ClusterAD [7, 8]. However, when the new data come in, the original model and the result cannot be adjusted accordingly. Cluster analysis needs to be conducted over again by re-entering a large amount of old data and new data, which is computationally intensive and unable to track the changes in clusters as flight operations evolve. In practice, airlines consolidate new flight data and perform Flight operations quality assurance (FOQA) or Flight Data Monitoring (FDM) analysis every month. In order to make the cluster-based anomaly detection method compatible with airline's practice, an online clustering method is needed to process new data every month.

In recent years, in the transportation direction, many methods have been developed to monitor the operation of transportation. [3, 7, 16, 17, 20, 22]. In the field of aviation, there have been many studies showing that clustering is a technology that can effectively identify various common patterns in operations. The Morning Reporting Package was one of the methods to detect anomalies in the aviation field at early stages [1]. The method uses statistical and mathematical based algorithms to identify abnormal flights, flight parameters, and flight phases. The Sequence Miner developed by Budalakoti is another method to detect anomalies. By inputting discrete flight data, such as binary switches inside the cockpit, the Sequence Miner algorithm can detect abnormalities in pilot switch operations based on Longest Common Subsequence (LCS) metric [2]. Srivastava et al. developed a statistical method that discretizes continuous data to combine discrete data with continuous data [19]. On top of this framework, Das et al. developed multi-core anomaly detection (MKAD). This method is based on the theory of multiple kernel learning and adopted one-class Support Vector Machine (SVM) to detect anomalies from a large set of continuous and discrete data [4]. Cluster-AD developed by Li et al. is a technique which is directly applicable to solve the anomaly detection problem for flight operations [7]. This method transforms time series data into a high-dimensional vector, which represents the trajectory of the takeoff phase or landing phase of a flight. After dimensionality reduction, the method adopts DBSCAN to cluster all the flights to identify the common operations. ClusterAD-DataSample uses the snapshot data at each time point as a data sample and adopts a Gaussian Mixture Model to automatically recognize multiple typical patterns of flight operations [8]. Melnyk et al. adopt a semi-Markov switching vector autoregressive (SMS-VAR) model to represent each flight and detect anomalies based on measuring the difference between the model's prediction and data observation [12]. A common challenge exists for all the above methods is that they all focus on the anomaly detection based on historical digital flight data. As new flight data come in, the methods cannot be updated rapidly.

Sato et al. developed an online EM algorithm for normalized Gaussian networks, which shows that the online EM algorithm can be seen as a stochastic approximation process to find the maximum likelihood estimator [15]. The idea of the online expectation-maximization (EM) algorithm for the method was derived by Xu, Jordan, & Hinton [21]. A number of studies adopt this idea to develop similar online EM algorithms for mixture models [6, 9, 14]. In this article, we adopt the idea of the online EM algorithm [21] to update the parameters of the Gaussian Mixture Model.

This study aims to develop an online cluster model to detect common patterns in the operations of aircraft based on flight data and update the identified patterns as new data come in. Compared with existing methods, the advantages of the new method lie in that it can (1) update the cluster model as new flight data are added into the model, and (2) track changes in clusters over time.

II. Method

The main purpose of the method is to perform cluster analysis and update the parameters of the cluster model as new flight data come in. The workflow of the method in this paper is illustrated in Fig.1, which consists of two parts: offline parts and online parts. For the offline part, the algorithm runs only once to get the initial parameters of the cluster model. For the online part, the algorithm runs every time when new flight data come in and the clusters are updated accordingly.

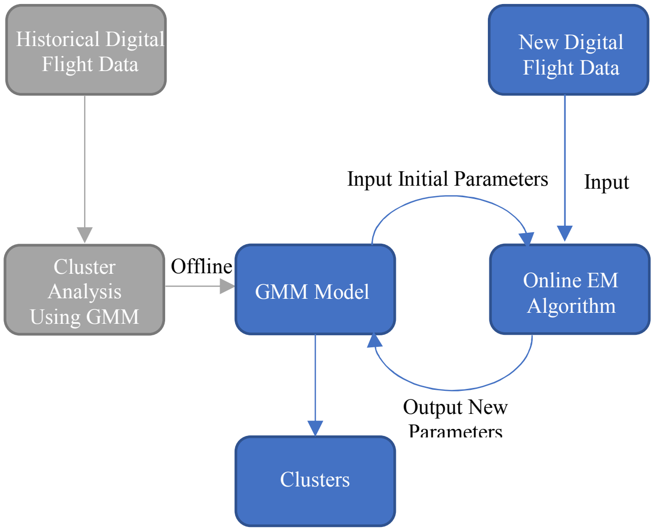

Fig. 1. Workflow of the method.

A. Data Transformation

Firstly, we need to normalize the fight parameters to have "zero mean and unit variance" for offline data. As for online data, we normalize them to the same standard with normalized offline data. After normalization, flight data are transformed into high-dimension vectors. Each vector $\mathbf{x}$ represents a single flight, including some selected parameters.

$$\mathbf{x} = [x_1^1, x_2^1, \ldots, x_f^1, \ldots, x_f^i, \ldots x_n^m] \tag{1}$$

where x_f^i is the value of the *f*th flight parameter at time *i*. n is the number of flight parameters and m represents the sample size for every flight parameter.

B. Cluster Analysis

After the data transformation, we develop a clustering method based on GMM to identify the different common patterns of the flight operations. The clustering method contains two part: offline GMM to obtain the initial parameters and online GMM to estimate the parameters based on new arrival data and update the initial parameters.

a) Offline Gaussian Mixture Model: We use clustering to identify the same pattern in flight data. The advantage is that the clustering algorithm can automatically assign similar vectors to the same cluster without the manually adding a label. The Gaussian mixture model is a typical clustering method which need to know the statistical properties of each cluster. The advantage is that parameters describe the characteristics of each Gaussian component, making it easy to update the parameters in the online algorithm. The GMM with the K component is given by:

$$p(\mathbf{x} \mid \lambda) = \sum\nolimits_{i=1}^{K} \omega_i g(\mathbf{x} \mid \mu_i, \Sigma_i) \tag{2}$$

where x is a set of M-dimensional vectors, $\lambda_i = \{\omega_i, \mu_i, \Sigma_i\}$ are the GMM parameters, K is the number of components of Gaussians and ω_i, i = 1,…, K, are mixture weights, satisfying $\sum\nolimits_i^K \omega_i = 1$, μ_i, i = 1,…, K, are the mean vectors and Σ_i, i = 1,…, K, are the covariance matrixes of Gaussians, and $g(\mathbf{x} \mid \mu_i, \Sigma_i)$ are the component Gaussian densities.

$$g(\mathbf{x} \mid \mu_i, \Sigma_i) = \frac{1}{\sqrt{(2\pi)^M |\Sigma_i|}} e^{-\frac{1}{2}(\mathbf{x}-\mu_i)' \Sigma_i^{-1} (\mathbf{x}-\mu_i)} \tag{3}$$

In order to determine the number of Gaussian mixture model components, K, sensitivity analysis was performed based on the Bayesian Information Criterion (BIC). The K with the smallest BIC is considered to be the best component.

The parameters of GMM ($\lambda^{initial}$) in offline part are obtained by using the expectation-maximization (EM) algorithm, which is a well-established method. We adopt the EM algorithm developed by Xu and Jordan and utilize the fact that at each iteration, the parameter increments have a positive projection on the gradient of the likelihood function (Xu et al. 1996).

In expectation step (E step), we need to calculated the log likelihood, $L_N(\mathbf{x} \mid \lambda)$. As for the maximization procedure, we need to update the parameters of GMM in each interaction to

The work was supported by the Hong Kong Research Grant Council Early Career Scheme (Project No. 21202716), and the National Natural Science Foundation of China Young Scientists Fund (Project No. 71601166).

get the convergence of likelihood. In the EM algorithm for GMM, the log likelihood is:

$$L_N(\mathbf{x} \mid \lambda) = \sum\nolimits_{j=1}^{N} \ln(\sum\nolimits_{i=1}^{K} \omega_i g(x_j \mid \mu_i, \Sigma_i)) \tag{4}$$

We can convert the EM algorithm to a gradient algorithm:

$$\lambda^{(t+1)} = \lambda^{(t)} + P^{(t)} L_N^{'}(\mathbf{x}, \lambda^{(t)}) \rightleftharpoons \lambda^{(t+1)} = \Xi(\lambda^{(t)}, \mathbf{x}) \tag{5}$$

where

$$P^{(t)} = \begin{pmatrix} P_{\omega}^{(t)} & 0 & 0 \\ 0 & P_{\mu}^{(t)} & 0 \\ 0 & 0 & P_{\Sigma}^{(t)} \end{pmatrix}$$

and

$$L_N^{'}(\mathbf{x}, \lambda^{(t)}) = {\partial L_N(\mathbf{x}, \lambda)} \Big/ {\partial \lambda} \Big|_{\lambda = \lambda^{(t)}}$$

Each GMM parameter is updated using the following equations.

$$\omega^{t+1} = \omega^{t} + P_{\omega}^{(t)} \frac{\partial L_N(\mathbf{x} \mid \lambda)}{\partial \omega} \Big|_{\omega = \omega^{t}} \tag{6}$$

$$\mu^{t+1} = \mu^{t} + P_{\mu}^{(t)} \frac{\partial L_N(\mathbf{x} \mid \lambda)}{\partial \mu} \Big|_{\mu = \mu^{t}} \tag{7}$$

$$\Sigma^{t+1} = \Sigma^{t} + P_{\Sigma}^{(t)} \frac{\partial L_N(\mathbf{x} \mid \lambda)}{\partial \Sigma} \Big|_{\Sigma = \Sigma^{t}} \tag{8}$$

where

$$P_{\omega}^{t} = \frac{1}{N} \left[\begin{pmatrix} \omega_1^{(t)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \omega_K^{(t)} \end{pmatrix} - \omega^{(t)} (\omega^{(t)})^T \right]$$

$$P_{\mu_i}^{(t)} = \frac{\Sigma_i^{(t)}}{\sum\nolimits_{i=1}^{N} \Pr^{(t)}(i \mid x_i, \lambda)}$$

$$P_{\Sigma_i}^{(t)} = \frac{2}{\sum\nolimits_{i=1}^{N} \Pr^{(t)}(i \mid x_i, \lambda)} \Sigma_i^{(t)} \otimes \Sigma_i^{(t)}$$

let

$$P_{\mu}^{(t)} = \begin{pmatrix} P_{\mu_1}^{(t)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & P_{\mu_K}^{(t)} \end{pmatrix}$$

$$P_{\Sigma}^{(t)} = \begin{pmatrix} P_{\Sigma_1}^{(t)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & P_{\Sigma_K}^{(t)} \end{pmatrix}$$

$\Pr(i \mid \mathbf{x}, \lambda)$ is the posteriori probability for each Gaussian component *i*:

$$\Pr(i \mid \mathbf{x}, \lambda) = \frac{\omega_i g(x|\mu_i, \Sigma_i)}{\Sigma_{j=1}^{K} \omega_j g(x|\mu_j, \Sigma_j)} \tag{9}$$

The proof of (6)-(8) can be given by the following the same line as the derivation of Theorem 1 of Xu and Jordan [21]. The GMM parameters are solved iteratively with the convergence criteria set as $L_N(x, \lambda^{t+1}) - L_N(x, \lambda^{t}) < \varepsilon$, where the termination tolerance (ε) is set as 1×10^{-6}. After the off-line GMM, we can get the initial parameters, $\lambda^{initial}$, based on the historical flight data.

b) *Online Gaussian Mixture Model:* In order to update the clusters as new flight data ($\mathbf{x}_{new}$), come in, we introduce a new algorithm to estimate and update the parameters of GMM based on new data and initial GMM parameters, namely online (recursive) EM algorithm. The online EM algorithm is composed of two parts. The first part is estimating a GMM (λ^{new}) based on newly arrived data. The second part is updating the previous GMM ($\lambda^{initial}$) based on the newly estimated GMM (λ^{new}) in order to obtain $\lambda^{updated}$.

Estimating a GMM of based on newly arrived data

In this part, we apply the EM algorithm to update the initial parameters when new data come in. We have:

$$\lambda^{(t+1)} = \Xi(\lambda^{(t)}, \mathbf{x}_{new}) \tag{10}$$

where $\lambda^{(t)}$ are parameters of *t*th interaction and $\lambda^{(t+1)}$ are parameters of *(t+1)*th interaction and Ξ is the parameter-updating function that we have mentioned above. To ensure and accelerate the convergence of the parameters, we set $\{a_t, t \geq 0\}$ as a sequence of positive numbers, satisfying

$$0 < a_t < 1, a_{t+1} < a_t, \lim_{t \to \infty} a \to 0$$

Then, at the *t*th iteration, the parameter λ is updated by:

$$\begin{aligned}\lambda^{(t+1)} &= (1-a_t)\lambda^{(t)} + a_t\Xi(\lambda^{(t)}, \mathbf{x}_{new}) \\ &= (1-a_t)\lambda^{(t)} + a_t[\lambda^{(t)} + P^{(t)}L_N^{'}(\mathbf{x}_{new}, \lambda^{(t)})]\end{aligned} \qquad (11)$$

For each interaction, the parameters of GMM can be updated by (11). If the likelihood converge is smaller than the termination tolerance ($\varepsilon = 1\times 10^{-6}$) the loop of interaction stops at *T*th interaction and output the updated parameter, λ^{T} , which is equal to λ^{new} in the next step.

Updating the previous GMM based on the newly estimated GMM

Now we need to combine the initial parameters obtained from the offline algorithm, $\lambda^{initial} = \{\omega^{initial}, \mu^{initial}, \Sigma^{initial}\}$, and the updated parameters $\lambda^{new} = \{\omega^{new}, \mu^{new}, \Sigma^{new}\}$ to get the final updated parameter, $\lambda^{updated} = \{\omega^{updated}, \mu^{updated}, \Sigma^{updated}\}$. Here we propose a straightforward way to update the GMM parameters:

$$\lambda^{updated} = (1-w)\lambda^{initial} + w\lambda^{new} \qquad (12)$$

TABLE I.

Pseudo code of online EM Algorithm

```
Input:
    Normalized vectors of new digital flight data $\mathbf{x}_{new}$
    Initial parameters of offline GMM, $\lambda^{initial}$
Output:
    Updated GMM parameters, $\lambda^{updated}$
Algorithm:
    1. Estimate a new set of GMM based on new data $\mathbf{x}_{new}$
        a) E step.
           Determine the log likelihood $L_N(\mathbf{x}_{new} | \lambda^{(t)})$ for
           $\mathbf{x}_{new}$ based on $\lambda^{(t)}$
        b) M step.
           for each Gaussian component:
               Update parameters using the following   equation.
               $\lambda^{(t+1)} = (1-a_t)\lambda^{(t)} + a_t[\lambda^{(t)} + P^{(t)}L_N^{'}(\mathbf{x}_{new}, \lambda^{(t)})]$
           end for
        c) Evaluate log likelihood:
           $L_N(\mathbf{x}_{new} | \lambda) = \sum_{j=1}^{N} \ln(\sum_{i=1}^{K} \omega_i g(x_j | \mu_i, \Sigma_i))$
           if (likelihood difference< $\varepsilon$ ):
               Output $\lambda^{new}$ ;
           else:
               $\lambda = \lambda^{(t+1)}$
               go to Step a.
    2. Combine initial parameters and updated parameters by the
       following scheme.
       $\lambda^{updated} = (1-w)\lambda^{initial} + w\lambda^{new}$
    3. return $\lambda^{updated}$
```

A new weighting parameter w is introduced here to balance the impact of new flight data versus historical flight data on GMM estimations. w is in the range of [0,1].

$$\omega^{updated} = (1-w)\omega^{initial} + w\omega^{new} \qquad (13)$$

$$\mu^{updated} = (1-w)\mu^{initial} + w\mu^{new} \qquad (14)$$

$$\begin{aligned}\Sigma^{updated} &= [(1-w)\Sigma^{initial} + w\Sigma^{new}] \\ &\quad + [(1-w)\mu^{initial}\mu^{initial^T} + w\mu^{new}\mu^{new^T}] \\ &\quad - \mu^{updated}\mu^{updated^T}\end{aligned} \qquad (15)$$

The value of w depends on the size of the offline dataset and the online datasets. The number of historical data is N as we mentioned above and we let N^{new} is the number of newly arrival data. Then, the value of w can be calculated by $w = \frac{N^{new}}{N + N^{new}}$. Finally, the Gaussian Mixture Model is updated to:

$$p(\mathbf{x} | \lambda^{updated}) = \sum_{i=1}^{K} \omega_i^{updated} g(\mathbf{x} | \mu_i^{updated}, \Sigma_i^{updated}) \qquad (16)$$

The pseudo code of the online EM algorithm is shown in Table I. After the updating of the parameters of GMM, we can use the new model to predict the cluster that each new data belongs to and output the new clusters and new parameters.

III. EXPERIMENT

A. Dataset

To demonstrate the performance of the proposed method, the algorithm was tested on a set of FDR data provided by an international airline, which contained landing phase of 1600 B777 flights. 256 flight parameters and 26 destination airports were included in the dataset. We set the beginning of the landing phase as 120 seconds before touchdown and convert position-related parameters into values relative to the airport location. 1600 flights were divided into two groups randomly: 1300 flights as training data and 300 flights as testing data, which is furtherly split into three groups equally.

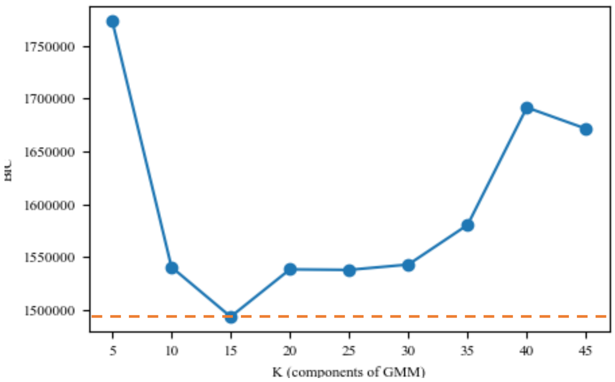

Fig. 2. The selection of component number based on BIC.

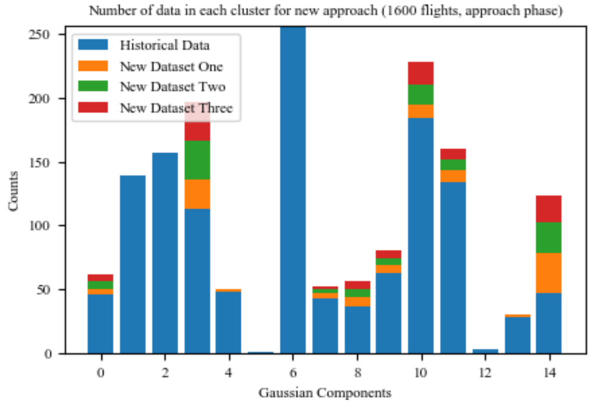

Fig. 3. Number of flights in each cluster in the new approach.

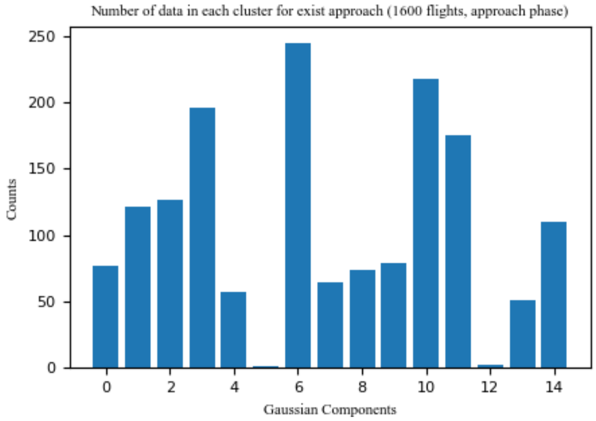

Fig. 4. Number of flights in each cluster in the existing approach.

B. *Offline Clustering*

We firstly used 1300 flights to build an offline Gaussian mixture model to get the initial parameters and clusters. Regarding the selection of K (number of mixture components), we found 15 to be the optimal value for this dataset as it gave the lowest BIC value, as shown in Fig. 2. Then, the parameters of the offline GMM can be obtained using the expectation-maximization (EM) algorithm.

C. *Online Clustering*

The purpose of offline clustering is to get the initial parameters and clusters of GMM. We have divided the original data into four groups, where the first one is offline dataset and the last three are online datasets. The online datasets represent three sets of data from three different months. Then we conducted the online cluster method on the three sets of stream data to update the parameters and clusters three times. At each time of update, we inputted one new dataset and the parameters that we got from the previous updating. Specifically, for the first new stream data, the input parameters are the parameters that we got from offline clustering. After three times of update, we got three different sets of parameters, which can be further analyzed in the next step.

D. *Results*

To evaluate of the proposed algorithm, we perform the cluster analysis on the data of all 1600 flights using the standard offline GMM algorithm, and then compare the results with the ones obtained by our adaptive approach. We let the parameters that we get by conducting the existing approach on all flight data are $\lambda^{all} = \{\omega^{all}, \mu^{all}, \Sigma^{all}\}$.

Fig. 3 and Fig. 4 shows the number of flights in each cluster by two approaches. The left figure shows the result of our approach, and the right one shows the result of the standard GMM with all data.

Fig. 5 shows the clustering results of our proposed approach using a visualization technique, T-distributed Stochastic Neighbor Embedding [10]. Each point in the figure represent the data of one flight. As shown in the figures, the three set of newly arrived data is progressively classified into existing clusters, and the cluster parameters are updated accordingly.

As for the parameters of the GMM in our new approach and the existing approach, we can see that the weights of our new approach (the yellow line in Fig. 6) and the weights for the existing approach (the blue line in Fig. 6) are similar to each other. The online algorithm didn't change the weights too much.

To compare the similarity between covariance matrix in our new approach and in the existing approach, we conduct a W statistic for each cluster. The null hypothesis H_0 for each cluster i is:

$$H_0 : \Sigma_i^{updated} \neq \Sigma_i^{all}, \ H_1 : \Sigma_i^{updated} = \Sigma_i^{all}$$

By calculating W value and p-value for each cluster, we collected all the results in Table II. Here we selected α=0.05 as significance level. As we can see, all the clusters except for the 5th cluster, 12th cluster and 13th cluster have the p-values that are smaller than 0.05. So we reject H_0 for the clusters except the three clusters above, and accept the hypothesis H_1 that for all the clusters except for the 5th, 12th and 13th clusters, the covariance matrixes of each cluster of our new approach is similar with that's of the existing approach.

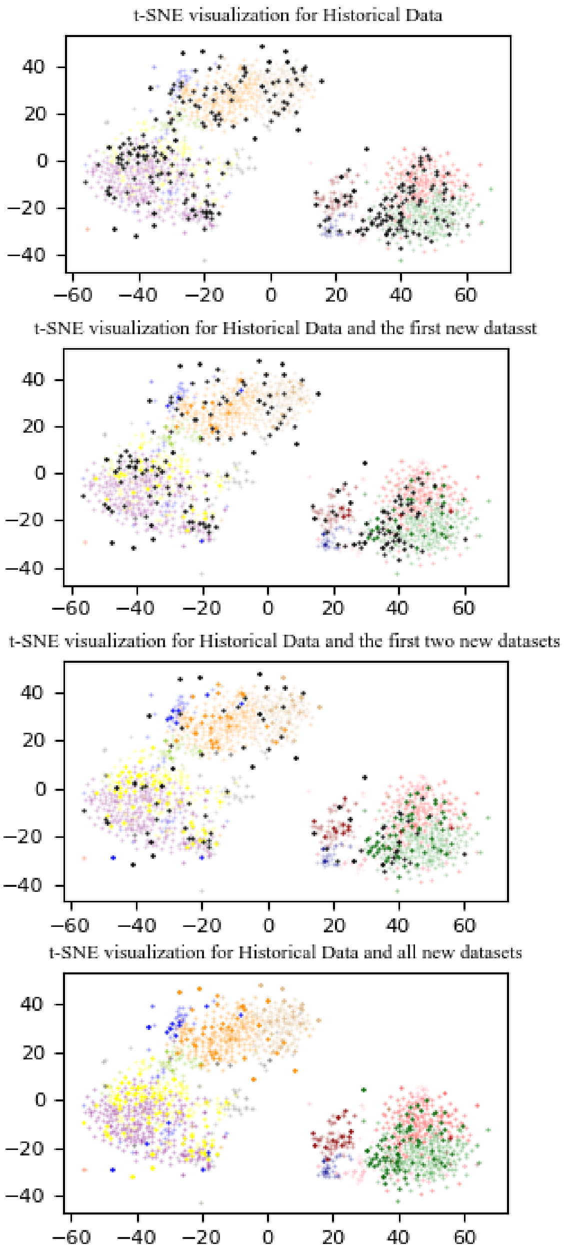

Fig. 5. Clustering results of our approach using t-SNE visualization technique

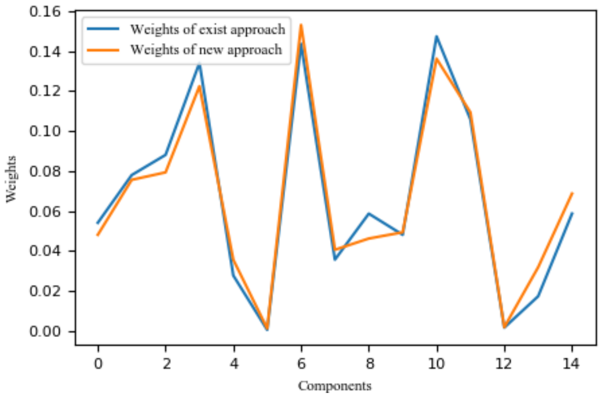

Fig. 6. Weights of the new approach and the existing approach

As for means and covariance of GMM model, the mean matrix is a 15×1920 matrix (15 Gaussian components and 120 seconds of data $\times$ 16 parameters of flights). To compare the similarity between means in our new approach and means in the existing approach, we conducted Hotelling's T^2 statistic to analyze the similarity between the means in our new approach and the existing approach. The null hypothesis H_0 for each cluster *i* is:

$$H_0 : \mu_i^{updated} \neq \mu_i^{all} \text{ , } H_1 : \mu_i^{updated} = \mu_i^{all}$$

By calculating the value of T^2 and p-value for each cluster, we also collected all the results in Table III. We also selected α=0.05 as significance level. Similar results were observed with covariance matrixes. Except for the 5th and 12th clusters, all the clusters have p-values less than 0.05. So we reject H_0 for the clusters except the 5th and 12th clusters, and accept the hypothesis H_1 that for all the clusters except the 5th and 12th clusters, the means in the new approach and the means in the existing approach are the same.

In summary, the proposed online algorithm can estimate GMM parameters relatively reliable compared to the standard offline approach. The exception of the 5th, 12th and 13th cluster may be caused by too few data in the clusters.

TABLE II.

W and p-value for covariance

Clusters	***W***	***p-value (α=0.05)***
0	0.914	< .05
1	0.641	< .05
2	0.745	< .05
3	0.966	< .05
4	0.755	< .05
5	0.500	0.496
6	0.832	< .05
7	0.796	< .05
8	0.868	< .05
9	0.761	< .05
10	0.900	< .05
11	0.684	< .05
12	0.377	1.000
13	0.505	0.366
14	0.679	< .05

TABLE III.

T^2 and p-value for means

Clusters	T^2	***p-value (α=0.05)***
0	75.524	< .05
1	11.507	< .05
2	52.617	< .05
3	410.930	< .05
4	101.140	< .05
5	3.7100	0.079
6	26.233	< .05
7	372.250	< .05
8	92.874	< .05
9	161.030	< .05
10	460.240	< .05
11	8.903	< .05
12	0.701	0.816
13	5.673	< .05
14	38.159	< .05

IV. Conclusion

Digital flight data are collected by airlines every month or even every day. It would be very time consuming to analyze all the data every time new data is generated. Even in the data reading phase, it takes a lot of time, which will greatly affect the efficiency of the airline operation. We developed a method for online clustering of new data, which can automatically update the original clusters when new data are added. The method was tested on real FDR data, which were provided by international airlines. Results show that this method is able to cluster new data and update the parameters of the clusters that already exist.

However, there are some limitations in our method. On the one hand, the method can only assign new data into the existing clusters and no more clusters could be generated. As airlines may adopt new technology or system on their aircraft, new operation patterns which are different with existing ones can be formed. On the other hand, new digital flight data are not the only data generated by the airlines. Two types of new data related to anomaly detection are generated at airlines every day. One is the onboard flight data; the other one is the airline safety experts' feedback on the anomaly detection results. Safety experts' feedback is as important as new digital flight data and is a good way for us to identify the realistic characteristics of each particular flight state. It can also help the airline to recognize different risks in various states. Therefore, a method that considers the real-time update of the two types of new data will be better to detect anomalies and provide more information to the airline.

In future steps, we will continue to complete the part of anomaly detection in some statistical ways to find outliers in our clusters and conduct some analysis on the outlier and clusters to identify different common patterns in operations. We will also improve our method to enable the updating of the number of clusters. It will be very useful to identify new clusters in new data, which will reveal more information for the airlines. Safety experts' feedback is another part that we are interested in. We plan to use the experts' feedback to identify the new outliers to avoid repeating work on the same type of outliers.

Acknowledgment

The work was supported by the Hong Kong Research Grant Council Early Career Scheme (Project No. 21202716), and the National Natural Science Foundation of China Young Scientists Fund (Project No. 71601166).

References

[1] Amidan, B.G. and T.A. Ferryman. Atypical event and typical pattern detection within complex systems. in 2005 IEEE Aerospace Conference. 2005. IEEE Big Sky, MT.

[2] Budalakoti, S., A.N. Srivastava, and R. Akella. Discovering atypical flights in sequences of discrete flight parameters. in Aerospace Conference, 2006 IEEE. 2006. IEEE.

[3] Chang, T.-H., et al., *Onboard measurement and warning module for irregular vehicle behavior.* IEEE Transactions on Intelligent Transportation Systems, 2008. **9**(3): p. 501-513.

[4] Das, S., et al. Multiple kernel learning for heterogeneous anomaly detection: algorithm and aviation safety case study. in Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. 2010. ACM.

[5] Eckstein, A. Automated flight track taxonomy for measuring benefits from performance based navigation. in Integrated Communications, Navigation and Surveillance Conference, 2009. ICNS'09. 2009. IEEE.

[6] Jordan, M.I. and R.A. Jacobs, *Hierarchical mixtures of experts and the EM algorithm.* Neural computation, 1994. **6**(2): p. 181-214.

[7] Li, L., et al., *Analysis of flight data using clustering techniques for detecting abnormal operations.* Journal of Aerospace information systems, 2015. **12**(9): p. 587-598.

[8] Li, L., et al., *Anomaly detection via a Gaussian Mixture Model for flight operation and safety monitoring.* Transportation Research Part C: Emerging Technologies, 2016. **64**: p. 45-57.

[9] Liu, Z., et al., *Online EM algorithm for mixture with application to internet traffic modeling.* Computational statistics & data analysis, 2006. **50**(4): p. 1052-1071.

[10] Maaten, L.v.d. and G. Hinton, Visualizing data using t-SNE. Journal of machine learning research, 2008. 9(Nov): p. 2579-2605.

[11] McLachlan, G.J. and K.E. Basford, *Mixture models: Inference and applications to clustering*. Vol. 84. 1988: Marcel Dekker.

[12] Melnyk, I., et al. Semi-Markov switching vector autoregressive model-based anomaly detection in aviation systems. in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. ACM.

[13] Neal, R.M. and G.E. Hinton, A view of the EM algorithm that justifies incremental, sparse, and other variants, in Learning in graphical models. 1998, Springer. p. 355-368.

[14] Nowlan, S.J., Soft competitive adaptation: Neural network learning algorithms based on fitting statistical mixtures. 1991.

[15] Sato, M.-A. and S. Ishii, *On-line EM algorithm for the normalized Gaussian network.* Neural computation, 2000. **12**(2): p. 407-432.

[16] Shi, Q. and M. Abdel-Aty, Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. Transportation Research Part C: Emerging Technologies, 2015. **58**: p. 380-394.

[17] Shichrur, R., A. Sarid, and N.Z. Ratzon, *Determining the sampling time frame for in-vehicle data recorder measurement in assessing drivers.* Transportation research part C: emerging technologies, 2014. **42**: p. 99-106.

[18] Song, M. and H. Wang. *Highly efficient incremental estimation of Gaussian mixture models for online data stream clustering.* in *Intelligent Computing: Theory and Applications III*. 2005. International Society for Optics and Photonics.

[19] Sriastava, A., Discovering system health anomalies using data mining techniques. 2005.

[20] Toledo, T., O. Musicant, and T. Lotan, *In-vehicle data recorders for monitoring and feedback on drivers' behavior.* Transportation Research Part C: Emerging Technologies, 2008. **16**(3): p. 320-331.

[21] Xu, L. and M.I. Jordan, *On convergence properties of the EM algorithm for Gaussian mixtures.* Neural computation, 1996. **8**(1): p. 129-151.

[22] Zhang, J., et al., *Data-driven intelligent transportation systems: A survey.* IEEE Transactions on Intelligent Transportation Systems, 2011. **12**(4): p. 1624-1639.