Pattern recognition of UAV flight data based on semi-supervised clustering

N Wang¹,³,⁴,⁵, Z SH Xu², SH W Sun¹,³, Y Liu¹,³,⁴,⁵

¹School of Management Science and Information Engineering, Jilin University of Finance and Economics, China
²Shenyang Woozoom Technology Co., Ltd, China
³Laboratory of Logistics Industry Economy and Intelligent Logistics, Jilin University of Finance and Economics, China
⁴Jilin Big Data Research Center for Business, Jilin University of Finance and Economics, China
⁵Jilin Province Key Laboratory of Fintech, Jilin University of Finance and Economics, China

Email: ctuwangnan@126.com

Abstract. UAV is an unmanned aerial vehicle controlled by a remote radio signal or a trajectory planning software carried itself. It is widely used in military, civil and scientific research fields. However, due to the lack of real-time decision-making ability, the UAV has high fault rate. The flight quality assessment of UAV and the construction of fault prediction model can be used for debugging and fault-removing to customer’s aircraft, and also to increase the added value of the civilian UAV products. Before building a fault prediction model, a very important step is to identify the pattern of sampled data. For each group of flight data, the efficiency and accuracy rate of manual quality evaluation and fault identification are not acceptable. Based on the UAV flight data accumulated in the big data platform of an UAV production company in Shenyang, Liaoning Province of China, this paper proposes a semi-supervised clustering technique to do automatic pattern recognition for the sampling points. According to the characteristics of UAV flight data, two different methods are designed to choose initial centroids. Meanwhile, we use the existing normal flight data to train distance thresholds to combine some clusters to eliminate the resulting error clustering. Real flight data or flight test data with manually added labels are used to run the proposed algorithms to verify the recognition results. The experimental results show that the proposed methods greatly improve the efficiency and accuracy of adding precise labels to the historical flight data and play a role in assisting the manual recognition of sampling points while strengthening the management and statistics.

1. Introduction

The unmanned aerial vehicle (UAV) has a significant improvement in its autonomous ability compared with the manned vehicle, and the ability to survive is greatly strengthened. The UAV can replace humans to complete the task in a bad environment and do not need to worry about the safety of the driver's life, and can perform the task of high risk[1,2,3,4]. So, UAV has been widely applied in military, civilian and scientific research fields. The widely application of UAV requires that it should have a high stability in addition to the efficient completion of the task, that is, the UAV is required to...
detect the health status of itself and take the corresponding scheme according to the current state [5]. Due to the lack of real-time decision-making ability, the UAV has high fault rate. Fault prediction is the core of UAV health management technology. The flight quality assessment of UAV and the construction of fault prediction model can be used for debugging and fault-removing to customer’s aircraft, and also to improve the added value of the civilian UAV products. The methods of fault diagnosis can be divided into model-based and model-free methods. When the model is not accurate, the diagnostic accuracy of the model-based diagnosis method is low, so the diagnosis method that does not rely on the model is getting more and more attention. The advantage of the model-free method is that it does not require the accurate model of the system and has strong self-adaption ability. Its defects are complex construction method making it not conducive to practical application [6]. The model-free methods can be classified into data driven, knowledge based and discrete event-based methods. UAV produces a large number of real-time data during the flight process. Using these data and machine learning algorithms, such as neural network-based methods [7,8,9,10], the fault prediction model of aircrafts can be constructed. However, the real-time data produced in the flight process are all unlabelled data, which is not to determine which sampling points in a flight data are fault-free points of normal flight and which are the fault points. So before building a fault prediction model, a very important step is to identify the pattern of the sampling data, which is also partially for improving the whole flight portrait. The interval of sampling points for UAV flight data is usually in milliseconds, for example, the data involved in this paper comes from the big data platform of a Shenyang UAV production company, with the data sampling interval of 2ms. For each group of flight data, the efficiency and accuracy rate of manual quality evaluation and fault identification are not acceptable.

All measurement variables have strong correlations, rather than independent characteristics. Therefore, we speculate that the pattern between the variables in normal flight will be different from that after fault happens. Running the proposed clustering methods on one group of flight data, the patterns of the sampling points are distinguished. In this algorithm, two different methods of choosing initial centroids are designed according to the characteristics of UAV flight data. Meanwhile, we use the existing normal flight data to train distance thresholds to combine some clusters to eliminate the resulting error clustering. We use real flight data or flight test data to test our methods. The experimental results show that the proposed methods greatly improve the efficiency and accuracy of adding precise labels to the historical flight data and play a role in assisting the manual recognition of sampling points while strengthening the management and statistics.

2. Data pre-processing

2.1. Feature selection

The UAV system consists of the controlled object (aircraft), the controller (generally including the measuring element, the sensor) and the actuator (steering engine, propeller). In addition, some other auxiliary positioning components (such as millimeter wave radar) may also be installed on the basis of the specific circumstances. The flight state of the aircraft is automatically determined by the flight control system according to the height or vertical speed of the aircraft. The flight mode of the aircraft can be controlled by the remote controller. The control volume of the remote controller channel is different under different flight modes. According to the flight mode and control volume or certain mission planning, the flight control system calculates the expectation data of location and speed. The controller calculates some control quantities based on these control expectations and current aircraft status and drives the actuator to control the behaviour of the important components, such as accelerometers, gyroscopes, magnetic compass and other ones, so that the aircraft will produce a certain mode of motion. The flight sensors mounted on these components measure these movements, including the pitch angle, the roll angle, the corresponding angular velocity and other variables, and return the calculated values, and the barometer returns the measured height. These values as actual values are compared with expectation values, and some control quantities for adjusting attitude are obtained, and then, the flight is corrected. Therefore, we can filter out the data generated during the
flight process, remove the intermediate variables in the calculation process, and retain the important feature variables.

The UAV flight control system adopts closed loop control. In different flight modes, the variables involved in the control and calculation are different. There are more than one flight mode switching in one flight data. Therefore, the unified characteristic variables cannot be used. For example, for multi axis UAV, in attitude mode, the index of speed is not involved in calculation and control. For the helicopter's unique mode of non-aileron, the indices of attitude, position and speed can be eliminated. Therefore, this paper runs clustering algorithm respectively on different flight modes.

2.2. Data normalization
After manual screening, we have retained a total of 51 characteristic variables for the flight data obtained by the big data platform, all of which are numerical data. Because the range gap between the feature variables is large and the value of each characteristic variable tends to be concentrated, the linear normalization function (Min-Max Scaling) is adopted to convert the original data value to the value in range of [-1, 1]:

2.3. Feature merging and feature weighting

2.3.1. Feature merging. Abrupt separation of three axis posture is one common fault which will produce the poor following values of Pitch, Roll and Yaw. According to this, we get the new combination features through the difference between the expectation value and the measure value of Pitch, Roll and Yaw.

2.3.2. Feature weighting. Another common fault is sensor failure. Data show that about 50% of UAV faults are caused by sensor failures [11]. When the fault is identified manually, it is found that the three-axis data change returned by the accelerometer is the most dominant fault data, which generally indicates that the accelerometer vibration suddenly becomes larger and is the most frequent fault type. Data normalization eliminates the influence of eigenvalues on distance calculation results, thus weakening the change of this mode. We dynamically weighted two sets of three axis data returned by the accelerometer (one set is the fixed data, the other is the shock absorption data), that is, the greater the three axis values of a certain sampling point being higher than its previous sampling point, the higher the weight value added to it.

The weight function is $x_\times xxxxxxxxx_xxxxxx_xxxxx$, in which, ‘xxxx’ represents Euclidean distance function, xxx is the vector composed of the two sets of three axis data. If there is a fault, the acceleration value will change greatly, and the corresponding weight of the point will be bigger. However, this does not distinguish between a real fault and a sharp change in the acceleration caused by the unsuitable operation of human operator, so that these will also be classified as fault data.

3. Algorithm design

3.1. Choose the number of clusters
For simplification, we only consider one kind of fault in flight process, or a most significant fault mode, not considering the mode of multiple fault superpositions for now. In fact, through the analysis of the flight data, we find that in the actual flight process, there are few fault superimpositions, even if there is, it is also such a situation that a component fault happens first, and before a short time of crashing, or in the moment of crashing, there may appear several other faults. After fault happens, there are two cases of crashing and non-crashing for the aircraft. We hope to identify the normal flight data points, the fault points (if crashing, also containing the crashing point), and the points after crashing (if crashing). Therefore, the selected number of clusters is 2 (not crashing) and 3 (crashing...
after faults happen), and the clusters and the number of clusters with the minimum error are output as the final clustering results.

For the flight data without fault, we pre-train cluster centroids distance threshold, and two clusters with the inter centroid distance less than the threshold are considered to be an error pattern recognition. For example, for a group of flight data without fault, all of the sampling points in theory should belong to the same cluster, but because the minimum clustering number is 2, the clustering results are not reasonable. At this time, the distance between the two cluster centroids is smaller than that of the fault-free points cluster and the fault points cluster and so, we will merge the two clusters.

3.2. Choose the initial centroids

We design two methods to choose the initial centroids and run the proposed clustering algorithm respectively:

(1) Random selection of $k$ sampling points as the initial centroids;
(2) The two adjacent sampling points with the maximum change amplitude of the three axis acceleration values are selected as the first two initial centroids, and the farthest sampling point from the set of the first two initial centroids is selected as the third centroid.

The third centroid in the method (2) determines which is farthest from set $S$ (here composed of the first two initial centroids) by solving the following optimization problems, namely the optimal function:

$$
\text{max}_{p \epsilon P - S \text{min}_{p' \epsilon S} D(p', p^t)}
$$

This optimal function represents: For each points $p^t$ ($1 \leq t \leq |P - S|$, $|P - S|$ represents the number of the points in set $P - S$), let $d_t$ be the smallest distance from $p^t$ to all points of $S$, $a^t$ is the farthest point from $S$ if $d_j = \text{max}_{1 \leq t \leq |A - S|}\{d_t\}$.

3.3. Flight data pattern discrimination algorithm based on k-means clustering

For the method of randomly selecting $k$ centroids, we run $k$-means clustering algorithm 10 times on each group of data and get the clustering result with the smallest clustering error. The clustering error is computed by SSE (Sum of Squared Errors), and the distance between sampling points is calculated by Euclidean distance formula. The flight data pattern discrimination algorithm based on $k$-means clustering is described as $\text{PattDiscrBasedonKMeans}$, in which we suppose there are $m$ kinds of flight modes.

**Algorithm.** $\text{PattDiscrBasedonKMeans}(\text{DataSet}, \text{GenerateCentMethod}, \omega)$:

- **#** Dataset is the data set to be clustered, GenerateCentMethod is the two proposed methods to get the initial centroids, and $\omega$ is the centroid distance threshold of merging fault-free data.
- **Do** feature merging and feature weighting for Dataset
- **Choose** feature variables for Dataset according to the flight mode: $\text{DataSet}_1 \sim \text{DataSet}_m$
- **For** $k$ in $\{2, 3\}$, **do**:
  - $\text{SSE}_k = 0$
  - **For** each $\text{DataSet}_i$ ($1 \leq i \leq m$):
    - $\text{originalCentroids}_i = \text{GenerateCentMethod}(\text{DataSet}_i, k)$
    - $\text{centroids}_i, \text{clusterAssment}_i, \text{SSE}_i = \text{kMeans}(\text{DataSet}_i, \text{originalCentroids}_i)$
      - **#** clusterAssment, is the classifying results for each point of $\text{DataSet}_i$, $\text{SSE}_i$ is the sum of squared errors for all points, centroids is the set of centroids of the resulted clusters
      - **#** Here if the method of randomly selecting $k$ centroids is adopted, it will output the clustering result with the smallest clustering error of running 10 times $k$-means clustering algorithm
    - Determine the two points in centroids, with the smallest distance, if the distance is smaller than $\omega$, merge the corresponding clusters, derive the centroid of the new cluster
and revise the value of $\text{SSE}_i$. Repeat this step until all points are merged to one cluster or the distance is not smaller than $\epsilon_0$.

$$\text{SSE}_k = \text{SSE}_k + \text{SSE}_i$$

Number the $m$ group of clustering results separately, combine them as the final clustering result of DataSet

Output the clustering results corresponding to the smaller values of $\text{SSE}_2$ and $\text{SSE}_3$

4. Experimental verification

4.1. Choosing data and evaluation index

4.1.1 Choosing experimental data. The experimental data of this paper come from the real flight data of the big data platform of an unmanned aerial vehicle (UAV) company in Shenyang City, Liaoning Province of China. After screening the characteristics of the original data, we chose 17 sets of fault flight data (199883 sampling points) and 29 groups of normal flight data (450996 sampling points).

The original flight data dimension obtained by the big data platform is very high, which contains all kinds of meaningless intermediate calculation results. We discussed with the company's technical employees involved in our research to select 51 eigenvalues at last. After choosing proper flight data and manually identifying the fault from large flight records, 17 sets of fault flight data were obtained. These data are all with manual labels, including the identification of normal points, fault occurrence point (if crashing happened, also including the crashing point), the points after crashing (if crashing happened). At the same time, it also includes the manual discrimination about the cause of the fault and the reason for crashing.

Table 1. Seventeen groups of fault flight data.

| Data No | Number of flight modes | Number of all sampling points | Number of predefined normal points | Number of predefined fault points (including crashing point) | Number of predefined after-crashing points |
|---------|------------------------|------------------------------|-----------------------------------|------------------------------------------------------------|------------------------------------------|
| 1       | 3                      | 13695                        | 9204                              | 4479                                                       | 12                                       |
| 2       | 4                      | 6244                         | 4911                              | 984                                                        | 349                                      |
| 3       | 2                      | 21535                        | 17849                             | 87                                                         | 3599                                     |
| 4       | 2                      | 1687                         | 776                               | 911                                                        | 0                                        |
| 5       | 2                      | 8034                         | 7665                              | 347                                                        | 22                                       |
| 6       | 2                      | 23601                        | 21875                             | 1715                                                       | 11                                       |
| 7       | 2                      | 5534                         | 4291                              | 121                                                        | 1122                                     |
| 8       | 2                      | 2815                         | 2676                              | 118                                                        | 21                                       |
| 9       | 4                      | 13681                        | 11656                             | 1949                                                       | 76                                       |
| 10      | 1                      | 9296                         | 9200                              | 96                                                         | 0                                        |
| 11      | 3                      | 9399                         | 8688                              | 97                                                         | 614                                      |
| 12      | 2                      | 9416                         | 9261                              | 155                                                        | 0                                        |
| 13      | 1                      | 3841                         | 3743                              | 98                                                         | 0                                        |
| 14      | 1                      | 8442                         | 7616                              | 394                                                        | 432                                      |
| 15      | 1                      | 12262                        | 12156                             | 106                                                        | 0                                        |
| 16      | 2                      | 28055                        | 25270                             | 2785                                                       | 0                                        |
| 17      | 2                      | 22346                        | 9783                              | 4418                                                       | 8145                                     |
4.1.2. Evaluation. The common clustering evaluation indexes are purity and F-value, and F-value is more commonly used. The calculation of F includes two indexes: Recall rate and Precision rate.

- \[ \text{precision}[i][j] = \frac{\text{the number of those points predefined as the } i\text{th class and classified to the } j\text{th cluster}}{\text{the number of points in the predefined } j\text{th class}} \]
- \[ \text{recall}[i][j] = \frac{\text{the number of those points predefined as the } i\text{th class and classified to the } j\text{th cluster}}{\text{the number of points in the predefined } i\text{th class}} \]
- \[ f[i][j] = \frac{2 \times \text{precision}[i][j] \times \text{recall}[i][j]}{\text{precision}[i][j] + \text{recall}[i][j]} \]

In this way, we get the F values between each predefined class and each derived cluster. We generate the weighted mean value of F obtained by the different initial centroids selection methods used in this paper.

\[ F\text{-measure} = \frac{\text{sum}(f[i][j] \times \text{the number of points in the predefined } i\text{th class})}{\text{the number of total points}} \]

The larger the value is, the better the clustering result is.

4.1.3. Experimental design. In this paper, we do experiments to 17 groups of fault flight data respectively. Two different methods are used to select the initial centroids, and the corresponding F values are calculated. At the same time, the accuracy of discrimination for the fault occurrence point is calculated.

In the course of the experiment, in accordance with the different flight mode, 29 groups of normal flight data are used to run the clustering algorithm with cluster number of 2. The distance between the two clusters of each flight mode is calculated, and the average value of the distances of the 29 sets of data is taken as the threshold for merging the clusters of the normal sampling points.

4.2. Results and analysis

For each group of flight data, we cluster it according to different modes, and then combine and manually correct the clustering results of different flight modes according to the time stamp of the sampling points.

In the clustering results obtained in a certain mode, the cluster composed of those points closer to starting time is classified as the normal class (fault free); the cluster composed of those points closer to ending time is classified as the after-crashing class (if the results are 3 clusters); the other points are classified as the fault class.

Thus, we get the final clustering result of a group of flight data. Then we calculate the values of indexes for the clustering results when running the clustering algorithms with different methods of generating initial centroids: Method1 and Method2, corresponding to the two methods in Section 3.2. The results are shown in Table 2 and Table 3.

| Data No | Number of classified normal points | Number of classified fault points | Number of classified after-crashing points | Number of generated clusters | Predefined number of clusters | F-Value | The recognition rate of fault points |
|---------|-----------------------------------|----------------------------------|------------------------------------------|----------------------------|------------------------------|---------|-------------------------------------|
| 1       | 9189                              | 4506                             | 0                                        | 2                          | 3                            | 0.89    | 87.65%                               |
| 2       | 4877                              | 1077                             | 290                                      | 3                          | 3                            | 0.88    | 84.75%                               |
| 3       | 17768                             | 299                              | 3468                                     | 3                          | 3                            | 0.90    | 86.66%                               |
| 4       | 767                               | 920                              | 0                                        | 2                          | 2                            | 0.84    | 83.51%                               |
| 5       | 7767                              | 267                              | 0                                        | 2                          | 3                            | 0.88    | 80.14%                               |
| 6       | 21878                             | 1723                             | 0                                        | 2                          | 3                            | 0.89    | 85.54%                               |
| 7       | 4317                              | 207                              | 1010                                     | 3                          | 3                            | 0.88    | 85.01%                               |
| 8       | 2689                              | 126                              | 0                                        | 2                          | 3                            | 0.85    | 39.58%                               |
Table 3. The values of indexes for the clustering results of Method 2.

| Data No | Number of classified normal points | Number of classified fault points | Number of classified after-crashing points | Number of generated clusters | Predefined number of clusters | F-Value | The recognition rate of fault points |
|---------|-----------------------------------|----------------------------------|------------------------------------------|----------------------------|-------------------------------|---------|-------------------------------------|
| 1       | 9200                              | 4495                             | 0                                        | 3                          | 3                             | 0.90    | 90.26%                             |
| 2       | 4967                              | 987                              | 290                                      | 3                          | 3                             | 0.88    | 87.83%                             |
| 3       | 17878                             | 289                              | 3368                                     | 3                          | 3                             | 0.90    | 89.23%                             |
| 4       | 767                               | 920                              | 0                                        | 2                          | 2                             | 0.88    | 87.50%                             |
| 5       | 7731                              | 287                              | 16                                       | 3                          | 3                             | 0.87    | 85.54%                             |
| 6       | 21879                             | 1703                             | 19                                       | 3                          | 3                             | 0.90    | 90.21%                             |
| 7       | 4274                              | 150                              | 1110                                     | 3                          | 3                             | 0.90    | 89.18%                             |
| 8       | 2695                              | 120                              | 0                                        | 2                          | 3                             | 0.88    | 75.13%                             |
| 9       | 11438                             | 2148                             | 95                                       | 3                          | 3                             | 0.90    | 89.50%                             |
| 10      | 9168                              | 128                              | 0                                        | 2                          | 2                             | 0.88    | 87.50%                             |
| 11      | 8696                              | 107                              | 596                                      | 3                          | 3                             | 0.87    | 87.04%                             |
| 12      | 9200                              | 216                              | 0                                        | 2                          | 2                             | 0.91    | 90.50%                             |
| 13      | 3679                              | 162                              | 0                                        | 2                          | 2                             | 0.91    | 90.50%                             |
| 14      | 7677                              | 315                              | 450                                      | 3                          | 3                             | 0.88    | 87.73%                             |
| 15      | 12111                             | 151                              | 0                                        | 2                          | 2                             | 0.90    | 89.50%                             |
| 16      | 25765                             | 2290                             | 0                                        | 2                          | 2                             | 0.91    | 90.50%                             |
| 17      | 9691                              | 3655                             | 9000                                     | 3                          | 3                             | 0.88    | 87.73%                             |
| Avg     |                                   |                                  |                                          |                            |                               | 0.89    | 87.96%                             |

The average recognition rates of the two methods to the fault points are 71.68% and 87.96%, respectively. But we find that both of the two methods have low recognition rates for the fault starting point and the crashing point. Only by using the second method, in two of the 17 sets of data the fault starting point are relatively accurately located (the error is 2 sampling points and 13 sampling points). The two methods have no more than 200 sampling points (0.4 seconds) for the forward or backward errors in the fault location, but in practical applications, this error is still somewhat large. The F values obtained by the two methods are both relatively large, which indicates that the effect of clustering is good, and it is found that most of the sampling points in consecutive time are in the same cluster. If after the automatic clustering, a certain manual correction is added, for example, to make the continuous sampling points belong to the same cluster, the recognition rate of the fault points will be greatly improved.
5. Conclusion
The big data platform of UAV enterprises accumulates a lot of real time data during the flight process of UAV. Using these data to warn the fault of aircraft during flight and timely adopting corresponding measures can reduce the loss for users caused by the crashing.

The first step to establish the early warning mechanism is to identify the types of faults that may happen in the aircraft, and to label the large number of historical flight data accurately.

In this paper, a semi-supervised clustering technique is used to distinguish normal sampling points and fault sampling points in a flight process, which greatly saves the time of manual recognition and improves the accuracy of recognition. We will further study from the following two aspects in future: (1) By improving the initial centroids selection, merging features and modifying distance formula, we can improve the clustering algorithm and further improve the accuracy of fault recognition. (2) Merging multiple groups of flight data of the same type of aircraft, finding the optimal number of clusters and identifying a variety of fault types.

Acknowledgments
This work is supported in part by NSFC under Grant No. 61702213, Jilin Provincial Department of Education “Thirteenth Five Year Plan” Science and Technology Development Plan under grant No. JJKH20180463KJ, Jilin Province Science and Technology Development Plan under Grant No.2018010137JC, the Open Project of Laboratory of Logistics Industry Economy and Intelligent Logistics under Grant No.201701.

References
[1] Wang H B, Ge X Y and Song Y Zh 2009 The development trend of UAV and forecast technology Winged Missiles Journal 11 46–52.
[2] Zou X F, He Q H, He J L 2006 Development status and related technologies of UAV Winged Missiles Journal 10 9–14.
[3] Liu Ch Y 2010 Development of UAV technology in foreign countries Ship Electronic Engineering 30(1) 19–23.
[4] Wallace L, Lucieer A and Watson C 2012 Development of a UAV-Li DAR System with Application to Forest Inventory Remote Sensing 4(6) 1519–1543.
[5] He M L 2015 Researches on fault diagnosis and fault-tolerant control for unmanned aerial vehicle Master Thesis of Guizhou University
[6] Shen Y, Ding S X and Haghani A 2012 A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process Journal of Process Control 22(9) 1567–1581.
[7] Hao T, Tang Y Zh, Ren Y Q 2008 Application of BP neural network in sensor fault diagnosis for flight control system Computer Measurement & Control 16(5) 613–615.
[8] Li M, Xu X D 2002 Powell neural network method for sensor fault detection Thermal and Power Engineering 17(1) 73-75.
[9] Gu L CH, Zhang Y Y, Qiu D 2002 Research on Sensor Failure Detection and Diagnosis Based on Neural Network Journal of xi’an jiao tong university 36(9) 959-962.
[10] Wang J P 2002 Fault diagnosis technology based on Neural Network Multisensor data fusion Mechanical science and technology 21(1) 127-130.
[11] Zeghlache S, Saigaa D Kara K 2016 Fault tolerant control based on neural network interval type-2 fuzzy sliding mode controller for octorotor UAV Frontiers of Computer Science 10(4) 657-672.