Aviation Safety Risk Analysis Based on Bayesian Network Modeling

Yuhang Wang, Xinkai Yue

Wuhan University of Science and Technology Wuhan, China

768495071@qq.com, 392923258@qq.com

Abstract. The occurrence of serious flight accidents not only brings huge economic losses to airlines but also poses a great threat to passengers' lives. The occurrence of serious flight accidents will not only bring great economic losses to airlines but also cause great threats to the lives of passengers. Therefore, in this paper, based on the flight records of different flight crews, flight routes, airports, and specific flight conditions, aircraft safety risk analysis is conducted to calculate and evaluate the risk propensity. Firstly, the data are preprocessed, and then the QAR data are clustered to extract the key data items affecting the safe flight of the aircraft; then the data are transformed to the attached data, and a hierarchical clustering model is established based on the K-means clustering algorithm, which classifies the factors affecting the safety quality of the aircraft flight into the three main aspects of environmental factors, crew maneuvering, and aircraft state, and the key parameters are weighted by the Random Forest algorithm. weights of the key parameters. Then, based on the 3 criterion, the outliers exceeding 3 were selected as the dependent variables leading to the occurrence of deviation; based on the three-layer model, a Bayesian neural network model based on the three-layer model, and then determine the Bayesian nodes; finally, calculate the Bayesian node probability, and then provide a reasonable quantitative description of the flight maneuver.

Keywords: K-means clustering, Bayesian neural networks, aviation security, random forests.

1. Introduction

Flight safety is the basis for the stable development of the civil aviation transportation industry. With the rapid development of China's civil aviation, the research on flight safety is becoming more and more important. The occurrence of serious flight accidents will not only bring huge economic losses to airlines, but also cause great threats to the lives of passengers. Therefore, it is necessary to focus on flight safety issues, strengthen aviation safety research, comprehensively utilize existing data to strengthen scientific management, and monitor and warn risks through targeted and systematic preventive and control measures, so as to reduce the chances of flight accidents [1].

At present, all large-scale public transportation aircraft registered in China are required to install QARs, which record the performance and status of aircraft equipment during the entire flight process, including operation parameters, external environmental parameters, pilot maneuvering-related parameters, and other hundreds of data related to the flight of the aircraft.

At this stage, the analysis of each overrun event by most airlines is limited to a few seconds before and after the event; and the correlation between each event in the monitoring system cannot be effectively analyzed and identified. Therefore, after the occurrence of the event to do statistics on the event itself, cannot analyze the potential risks that led to the occurrence of the event, and cannot predict the occurrence of the event through the occurrence of risk factors. For hidden events that do not occur but are very close to exceeding the limit, it is difficult for flight quality monitoring to obtain this information, so that the pilot who triggered the event may be at the edge of danger [2].

Therefore, based on the flight records of different flight crews, flight routes, airports, and specific flight conditions, modeling and analyzing the data to calculate and evaluate the risk propensity, so as to carry out targeted safety management, investigate potential safety hazards, and improve the safety performance is an effective measure to strengthen the flight safety issue.

2. Flight safety-critical data extraction

2.1. Data preprocessing

Before extracting key parameters from QAR data, we need to preprocess the data. The work of data preprocessing is divided into two parts. First, the collection frequency of QAR data is unified. Among the different indicators of QAR data, the frequency of indicators is different due to the different collection equipment and the different degree of attention to different indicators. In order to simplify the calculation process, we take once a second as the harmonized frequency. By finding the mean value of the variable every second to be used as the parameter value.

For example c1, c2 ,…… , c_n is the empty attribute value in the column of QAR data, if c1 , cn is empty, we use the first non-empty data in c2 ,…… , c_{n-1} to fill in, if it does not exist, fill in 0, and similarly, use the first non-empty value in c2 ,…… , c_{n-1} to fill in c_n; if c_i is not the boundary of the attribute value, we take the average of the closest attribute values of ci on both sides of c_i . If ci is not its boundary, then take the average of the closest attribute values on both sides of ci as the value of c_i. The formula is as follows:

$$c_i = c_{i+1}(1 \le i \le r) \tag{1}$$

$$c_{i+(i+j)/2} = (c_i + c_j)/2(r \ge i, j \le q) \tag{2}$$

$$c_i = c_{i-1}(q < i \le r) \tag{3}$$

Where c_r , c_q is not empty, otherwise fill in 0.

Due to the large number of indicators in the QAR data, the clustering results of the variables will be affected by the different scales or orders of magnitude of the indicator variables. In order to make the clustering unaffected by the order of magnitude, it is often necessary to do the transformation. The formulas for the commonly used transformation methods are as follows:

$$\bar{x_k} = \frac{1}{n}\sum_{i=1}^{n} x_{ik} \tag{4}$$

$$\sigma_k^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_{ik} - \bar{x_k})^2 \tag{5}$$

$$x_{ik} = \frac{x_{ik} - x_k}{\sigma_k} \tag{6}$$

Where $\overline{x_k}$ is the mean of the Kth variable, $\overline{\sigma_k}$ is the standard deviation of the Kth variable, and x_{ik} is the i-th data of the Kth scalar after normalization.

By this transformation, we transform the different scalar indicators into parameters with similar values. The processed data are shown in Table 1, which is a partial truncation of the metrics obtained after the above transformation of the QAR data. It can be seen that the parameter values between all the indicators are relatively close to each other, which is convenient for the subsequent clustering operation. In the process of calculating using the above formula, due to the fact that some of the QAR parameters take the value of the binary value of the state change, the standard calculation result is 0, and these parameters have little significance to our parameter extraction and will not be dealt with.

2.2. Segmentation of flight stages based on K-means clustering algorithm

Overrun events come from the flight quality monitoring system. We can understand overrun events as risks, but the risks encountered by aircraft during flight are not fully covered by the flight monitoring system. The causes of unsafe events are complex and the reasons for their formation are different. Some unsafe events are the result of a combination of factors, while others are accidental occurrences. Insecurity is caused by risk, which can be quantified using QAR data. However, due to the large volume of QAR data and the large number of data items, a relatively large amount of the data is of low practical value and contains redundant information. So, we need to extract the key

parameters from the QAR data and then analyze the importance of the key parameters to find out the decisive parameters that lead to the risk.

The purpose of clustering in this session is to obtain clusters with similar changes in the indicators. It is inaccurate to measure the aircraft takeoff, flight and landing process by considering only individual parameters, for example, the aircraft takeoff and landing levers are turned off during the ascent of the takeoff and turned on before the landing, so much so that it is necessary to cluster the QAR data for a better evaluation of the flight quality of the aircraft during the flight process.

The K-means algorithm is a classical algorithm for clustering algorithms, and its clustering step is a cyclic iterative algorithm as follows:

(1) Assuming that we want to do clustering on N sample observations, and require clustering as K classes, first select K points as the initial center point.

(2) Next, divide all observations into the classes in which each centroid is located according to the principle of minimum distance from the initial centroid.

(3) With a number of observations in each class, calculate the mean of all sample points in the K classes as the K centroids for the second iteration.

(4) Steps 2 and 3 are then repeated based on this center until convergence (center points no longer change or the specified number of iterations is reached) and the clustering process is complete.

The k-means algorithm is to divide the data into different clusters, the goal is to have small differences in the same cluster and large differences between different clusters, the sum of error squares can be used as the objective function, the formula is as follows:

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 \tag{7}$$

Where C_i is the first class, p is all sample points in C_i, is the center of mass of C_i (the mean of all samples in C_i), and SSE is the clustering error of all samples, which represents the clustering effect.

In order to find the K-value in step 1 of the algorithm, the elbow method will be used. The core idea of the elbow method is as the number of clusters K increases, the samples will be divided more finely, and the degree of aggregation of each cluster will be gradually increased, so the sum of squared errors and SSE will naturally become smaller. When K is smaller than the true number of clusters, the increase of K will greatly increase the degree of aggregation of each cluster, so the decline of SSE will be very large, and when K reaches the true number of clusters, and then increase the degree of aggregation of K returns will be rapidly become smaller, so the decline of SSE will be reduced abruptly, and then with the K value continues to increase and tend to flatten out, which means that the relationship between the SSE and K is the shape of an elbow, and the elbow corresponds to the K value. The K value corresponding to this elbow is the true number of clusters of the data. The optimal number of clusters K value can be calculated after bringing the data, the specific results are shown in Figure 1 below.

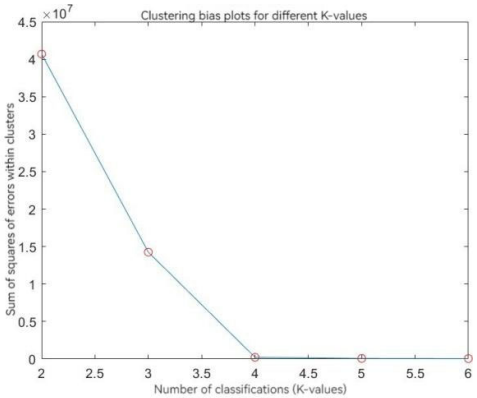

Fig. 1 K-means clustering result plot.

From Figure 1, it can be seen that at the classification number K=3, the curve forms an inflection point, and the change of its slope is the largest, which indicates that the flight phases of the airplane should be divided into 3 phases.

2.3. Building a hierarchical clustering model

When R-type clustering is performed, variables with large differences in change trends are separated and those with small differences are clustered together. Variables with high similarity are clustered together first. After the clustering process is completed, the number of variables is not reduced, only clustering between variables occurs. Therefore, it is also necessary to approximate the variables within the group to determine the best categorization between variables. For the selection of the clustering distance, we choose the Euclidean distance, which is calculated by the following formula:

$$d = \sqrt{\sum_{i=1}^{N} (x_{mi} - x_{ni})^2} \tag{8}$$

Where d represents the Euclidean distance between two indicators, x_{mi} represents the value of the ith parameter for the mth indicator and x_{ni} represents the value of the ith parameter for the nth indicator. After calculating the distance between the two indicators, we form a similarity matrix between the variables.

Through many iterations, we finally classify the parameters affecting the flight safety quality of the airplane into three aspects, and the key parameters selected for each aspect are:

① Environmental factors: magnetic heading, radio altitude, and aircraft weight.

② Crew maneuvering: wind direction, attitude, down track deviation, heading track deviation, calculated airspeed, low speed.

③ Aircraft status: descent rate, wind speed, G-value average, stick volume, pitch rate, slope, disk volume.

2.4. Modeling safety factor weights based on the random forest algorithm

The random forest algorithm is a machine learning algorithm. In the process of generating many decision trees, by randomly sampling the sample observations and feature variables of the modeling dataset respectively, the result of each sampling is a tree, and each tree generates rules and classification results in line with its own attributes, while the forest eventually integrates the rules and classification results of all decision trees to realize the classification of the random forest algorithm. The flowchart of the random forest algorithm is shown in Figure 2.

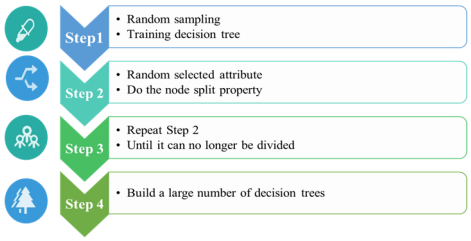

Fig. 2 Random Forest Algorithm Flowchart

The training rule for random forest is as follows:

➢ Step1: If the size of the training set is N, for each tree, randomly and with a putback, draw N training samples from the training set (this sampling method is called bootstrap sample method) as the training set for that tree.

➢ Step 2: If the feature dimension of each sample is M, specify a constant m<<M, randomly select a subset of m features from the M features, and each time the tree is split, select the optimal one from these m features.

➢ Step 3: Each tree is grown to the maximum extent possible and there is no pruning process.

Finally, the weights of each key parameter are calculated, in order of their importance: aircraft weight (31%), wind speed (17%), radio altitude (15%), wind direction (7%), and ground speed (5%), etc., as shown in Figure 3.

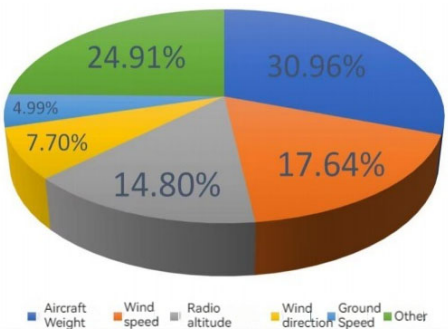

Fig. 3 Key Parameter Weighting Chart

3. Quantitative description of flight maneuvers

3.1. Extraction of outliers based on the 3σ criterion

The 3σ criterion is based on repeated measurements with equal precision that are normally distributed and cause interference or noise in the singular data that makes it difficult to satisfy the normal distribution. If the absolute value of the residual error $v_i > 3\sigma$ for a measurement value in a set of measurement data, the measurement value is bad and should be rejected. Usually, the error equal to $\pm 3\sigma$ is regarded as the limiting error. For normally distributed random errors, the probability of falling outside $\pm 3\sigma$ is only 0.27%, and it is very unlikely to occur in a finite number of measurements, therefore, there exists the 3σ criterion. The 3σ criterion is the most commonly used and simplest criterion for discriminating the gross errors, and it is usually applied to the case when the number of measurements is sufficiently large ($n \geq 30$) or when $n > 10$ to make the gross discriminations.

$$P(|x - \mu| > 3\sigma) \leq 0.003 \tag{9}$$

Finally, by filtering to find outliers outside of 3σ, we extracted 7 key parameters affecting the G-value of the airplane, as shown in Table 1 below.

Tab. 1 Table of factors affecting G-value

Serial No.	Factor
1	Altitude
2	Left/Right Hair Throttle Lever Factors
3	Groundspeed
4	Cross-examination
5	Pitch angle ratio
6	Left/right hairpin throttle position
7	RUDD Location
8	Stroke volume

3.2. Bayesian network-based risk analysis modeling

Flight operation is a dynamic and complex system with many factors acting together, in Problem 1 we extracted the risk factors affecting flight safety, the occurrence of unsafe events is often the result of the joint action of multiple factors, and there is a certain logical connection between different risks. Therefore, in the analysis of flight risk and target events in the landing phase, we use the Bayesian network to conduct inference and diagnostic analysis.

The core content of the Bayesian principle is the conditional probability, P(A|B) denotes the probability of event A occurring if event B occurs:

$$P(A|B) = P(AB)/P(B) \tag{10}$$

The role of Bayes' theorem is reflected in the reverse derivation: usually, we can directly calculate P(A|B) in the data, P(B|A) is difficult to obtain directly, but we are more concerned about P(B|A), Bayes' theorem will open the way for us to obtain P(B|A) from P(A|B).

$$P(B|A) = P(A|B)P(B)/P(A) \tag{11}$$

Bayesian networks are an extension of Bayes' theorem, Bayesian networks are a mathematical model based on probabilistic reasoning, usually presented as a graphical network. A Bayesian network takes the form of a directed acyclic graph, where the nodes in the graph represent different variables and the links between the nodes represent the correlations between the variables. The nodes in a Bayesian network can be an abstraction of any problem or attribute, for example, the age of the aircraft, airport elevation, wind direction, and other factors can be used as nodes, the value of the stage can be discrete, continuous, and other different distribution states; the correlation between the nodes through the connecting lines qualitatively, the strength of the relationship is expressed through the conditional probability of the root node in the network, through the data computation or the work of the prior probability assigned to the root node. It is a basic principle that no ring structure is allowed in Bayesian networks.

We use the following process to build the Bayesian network structure. The first step is to determine the scope of the problem, there are many stages of flight operation, and the external environment, flight status, and other factors are different in each stage, if we want to do quantitative risk analysis based on QAR data, we can only choose one flight stage, and the landing stage is the stage with the highest risk coefficient of flight operation, so we choose the landing to study the relationship between the risk factors and the target event. After determining the scope of the problem, the second step is the study of civil aviation industry knowledge to sort out, able to qualitatively analyze the landing phase environment, risk, and personnel operations. Indicators affecting pilots were identified, and then, based on QAR data, we quantified the indicators. The third step is to build the structure of the Bayesian network based on the quantified indicators, based on a large amount of historical QAR data, to determine the prior probability and conditional probability of the nodes in the network; and finally, through the already established analytical model, to carry out the inference and diagnostic analysis of the risk and the current events.

3.2.1 Modeling the three layers

After extracting the outliers through the 3σ criterion, a three-layer model was built to obtain the stratification results shown in Figure 4 below.

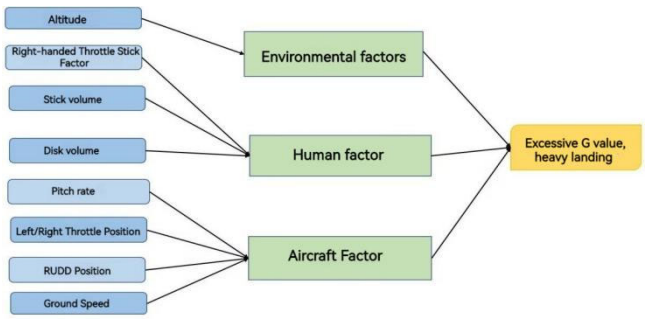

Fig. 4 Three-layer model result chart

3.2.2 Building Bayesian networks

A Bayesian network is a directed acyclic graph consisting of nodes representing variables and directed edges connecting these nodes. The nodes represent random variables and the directed edges between the nodes represent the mutual relationships between the nodes (pointing from the parent

node to its children), expressing the dependencies between the variables in terms of conditional probabilities, and the information in terms of prior probabilities for those that do not have a parent node [3].

Let G be a Bayesian node defined on $\{X_1, X_2, \cdots, X_N\}$ a Bayesian network whose joint probability distribution can be expressed as the product of the conditional probability distributions of the individual nodes:

$$P(X) = \prod_i p_i(X_i|Par_G(X_i)) \tag{12}$$

Where $Par_G(X_i)$ is the parent of node X_i, and $p_i(X_i|Par_G(X_i))$ is the conditional probability table of the node.

3.2.3 Determination of Bayesian nodes

The first stage of Bayesian network building is to identify the nodes in the network and the possible states of each node. Combined with the situational awareness in the landing phase, we identify the nodes in the network layer by layer. In the risk characterization layer, the state of the aircraft, human factors, and the external environment form an interactive whole that ultimately affects the safety of the aircraft. We identify the following parameters as nodes in the risk characterization layer, and we abstract the distribution of the nodes into discrete binary or ternary states for ease of computation. For example, the values of "aircraft approach height" are "high, middle, low", and the values of "pilot training status" are "good, poor".

The final Bayesian nodes are altitude, right throttle stick factor, left/right throttle position, average G-value, pitch rate, RUDD position, and ground speed [4].

3.2.4 Bayesian node probability determination

Finally, the Bayesian node probabilities are obtained by calculation as shown in Table 2 below, which shows that the deviation factors leading to one loose stick maneuver are mainly caused by altitude, right hair throttle stick factor, ground speed, stick volume, disk volume, pitch rate, left/right hair throttle position and RUDD position, and their prior probabilities are 0.26, 0.15, 0.25, 0.24, 0.05, 0.05, respectively, 0.05, and 0.05, respectively, indicating that the factors leading to a significant decrease in the G-value are mainly altitude, ground speed, and rod volume [5].

Tab. 2 Table of node probabilities

Node Name	State of affairs	Distributed probability value	Prior probability
Altitude	High	0.07	0.26
	Middle	0.08	
	Low	0.11	
Right hair throttle lever factor	Yes	0.05	0.15
	No	0.10	
Groundspeed	High	0.08	0.25
	Middle	0.02	
	Low	0.05	
Stroke volume	High	0.19	0.24
	Middle	0.03	
	Low	0.02	
Cross-examination	Yes	0.03	0.05
	No	0.02	
Pitch angle ratio	Yes	0.045	0.05
	No	0.005	
Left/right hairpin throttle position	Yes	0.04	0.05
	No	0.01	
RUDD Location	Yes	0.03	0.05
	No	0.02	

4. Summary

The random forest algorithm used in this paper, the algorithm can get high dimensional (many features) data, and does not need to reduce the dimensionality, nor need to do the feature selection. It can also determine the importance of features, and the interaction between different features, and the training speed is relatively fast, easy to make parallel methods. The K-means cluster analysis algorithm used in this paper is a classical algorithm for solving clustering problems, which is simple and fast. For this paper, the clustering effect is better when the clusters of chemical components are dense and the difference between clusters is obvious.

For K-means clustering method can be improved, for the K-means algorithm's difference measure principle, proposed a clustering method based on the improvement of the weight value, differentiate between different dimensions of the data, realize the different dimensions of the attributes of the clustering results of the different degree of influence, can effectively improve the accuracy of the clustering, but due to the full-weight value is not easy to obtain, so there is no improvement made in this model. Then we chose the Bayesian classifier algorithm, but for the small number of samples given in the data, the accuracy rate of choosing the decision tree classifier is relatively high, and the most important judgmental factors are well arranged in the position close to the root of the tree, which is more intuitive to represent the results, but due to the technical reasons failed to realize the decision tree classifier algorithm, so instead of using the Bayesian classifier algorithm.

References

[1] Liu Liu. Research on flight risk in landing phase based on QAR data [D]. Chongqing:Chongqing University,2018.

[2] Shen Xiaofeng,Yu Binhua,Qiao Hui. Research on airline QAR data mining and application [J]. Journal of Civil Aviation,2019,3(5):28-31.

[3] FENG Junjun,HE Xiaochun,WANG Haipei. Research on microblog topic tracking technology based on plain Bayesian network [J]. Computer and Digital Engineering,2017,45(11):2244-2247,2284.

[4] Deng Xiaoyong,Zhang Dongyang,Jin Keke. Aircraft vulnerability assessment method based on fuzzy Bayesian network [J]. Firepower and Command and Control,2023,48(1):92-98.

[5] YE Ziyang,CHEN Yanyi,ZHANG Peilin,et al. Analysis of inland vessel traffic accidents based on data-driven Bayesian network [J]. Safety and Environmental Engineering,2022,29(1):47-57.