Research and evaluation of attribute reduction method based on flight parameter data

Wenbing Chang, Zhenzhong Xu and Shenghan Zhou

School of Reliability and System Engineering, Beihang University, Beijing, China

1Email: zhoush@buaa.edu.cn

Abstract. The landing phase of an aircraft is a frequent phase of a flight accident. To ensure the safety of aircraft landing is the focus of the aviation field and strive to overcome the problem. With the advent of the data age, landing safety research based on flight parameters is the research trend in the aviation field. At present, there are few studies on landing safety based on the data of flight parameters. In this paper, based on the neighborhood rough set, we use the forward greedy algorithm to compare the attribute reduction with the factor analysis, and use the random forest classification model to evaluate and analyze.

1. Introduction

Aircraft landing process is more complex, vulnerable to aircraft equipment reliability, driver experience and status, the landing environment and other factors, resulting in multiple aircraft landing phase accident. With the installation of QAR (Quick Access Recorder) equipment, the study of aircraft landing quality has shifted from the traditional driver model evaluation to data-driven research. Cross et al used the neural network method to verify the correlation between the flight data and the landing load [1]. Lane et al. used QAR data to simulate non-linear flight dynamics of twin jet aircrafts. The results show that the vertical air flow changes are not conducive to the descending speed control, resulting in greater landing load [2]. Based on the flight data of QAR records and the principle of aircraft heavy landing, Wang Lei et al. constructed a landing risk assessment model represented by heavy landing [3]. With the development of QAR technology, flight parameters recorded by QAR have increased significantly, among which there are missing values, abnormal values, noise values, and redundant information. Neighborhood rough sets are suitable for solving these ambiguities and uncertainties. It is also suitable for numerical data processing. The initial cleaning of the original data and the attribute reduction of the flight parameters data using the neighborhood rough set can eliminate unnecessary parameters and improve the overall quality of the flight parameters. The aircraft landing quality prediction model built by using flight data was to provide guidance for the landing of the aircraft and ensure the safety of the aircraft landing. This paper presents a set of attribute reduction methods for flight parameters based on neighborhood rough sets. This paper proposes a value $\lambda$ to regulate the neighborhood radius and explore the effect of neighborhood radius on the reduction effect. In order to reduce the number of computations in the positive domain, this paper uses a combination of the neighborhood rough set and the forward greedy algorithm to achieve rapid reduction of flight parameters. Compared to factor analysis, this way works well.
2. Methods and models

Attribute reduction for multi-parameter is the core of rough set theory. Rough set theory was first discovered by Pawalak in 1982 as an important data mining tool, mainly used to solve the problem of ambiguity and uncertainty [4]. The traditional rough set can't directly deal with the shortcomings of continuous data so that its application in many areas is limited. To improve the scope of rough sets, Wang [5] Khan M I [6], Yazdani M [7], Xu F [8], Gul M [9] made related improvements.

The neighborhood rough set model uses the $\delta$ - neighborhood relation of the distance measure function instead of the equivalence relation in the traditional rough set. This method is suitable for direct processing of numerical data.

**Definition 1 $\delta$ - neighborhood** [10]: a non-empty finite set $U=\{x_1,x_2,\ldots,x_n\}$ on a known $n$ dimensional real space $\Omega$, assuming that $N=\{U,\varphi\}$ is a neighborhood, and $\forall x_i \subseteq U$ for $\delta$ - neighborhood is defined as:

$$\delta(x_i) = \{x \mid \varphi(x_i,x) \leq \delta, x \in U\}$$  \hspace{1cm} (1)

Although the attribute reduction of neighborhood rough set model has been successfully implemented in many applications, such as feature selection, classifier, rule algorithm and so on. It is still inefficient to calculate the attribute reduction, especially calculating the neighbor of each sample. Based on the above shortcomings, this paper uses attribute reduction algorithm based on the forward greedy algorithm. The greedy search algorithm needs only to divide the condition attributes once, form a set of neighborhood relation matrix, and then simplify the neighborhood relation matrix. This can improve algorithm efficiency.

The forward greedy attribute reduction algorithm based on the neighborhood rough set is different from the traditional rough set algorithm in that it introduces the neighborhood radius and the lower threshold of importance degree to realize the information grain of the numerical data and control the convergence speed of the reduction process.

The steps are as follows:

- **Input**: Neighbor decision system and lower limit of importance degree (Control the speed of convergence).
- **Output**: Reduction set.

(1) Standardization of domain attributes.
(2) Calculate the standard deviation of each condition attribute.
(3) Set parameter $\lambda$ to calculate the radius of each attribute field.
(4) Initialize the reduction set.
(5) Calculate the positive domain.
(6) Calculate the importance.
(7) If the importance is greater than the lower limit of importance degree, output the reduction set. Otherwise, go to (3).

Random forest is a commonly used machine learning algorithm proposed by Leo Breiman [11] in a paper published in 2001. Random forest is a combination of classifier, which is composed of multiple decision trees, and finally by simple voting method to determine the final classification results. Tetschke F [12], Zhao [13], Wu [14], Ding [15], Zhu [16] did research in related fields.

The calculation of random forest classification is as follows:

- First, using $N$ to denote the number of original samples, and $M$ to denote the number of features.
- Using Bootstrap to randomly select $k$ sample subsets of which the size is $N$ from the original sample data set. Construct $k$ decision tree classification models. The sample set that was not selected every time in the original sample set constitutes an out-of-bag (OOB) data set.
- Each selected subset of samples is constructed as a decision tree model. At each node of the tree, $m$ features ($m \leq M$) are randomly selected from the $M$ feature variables. According to
the maximum principle of node purity, select one feature from these \( m \) feature variables for node segmentation.

- No branches need to be pruned for each decision tree to maximize its growth.
- Random forest is composed of so many decision trees, according to the voting results of each decision tree to get the random forest classification results.
- The out-of-bag data (OOB) generated from each sampling was used to predict the classification accuracy of the random forest model.

In order to analyze the classification effect of random forest classification and prediction model more clearly and stochastically, this paper constructs the following forecasting results of landing quality: True High Landing Quality (THQ), Flase High Landing Quality (FHQ), True Low Landing Quality (TLQ), Flase Low Landing Quality (FLQ). Model classification predictions as shown in table 1.

| Record indicators | True situation     | Result             |
|-------------------|--------------------|--------------------|
| THQ               | Landing high quality| Landing high quality|
| FHQ               | Landing low quality | Landing high quality|
| TLQ               | Landing low quality | Landing low quality |
| FLQ               | Landing high quality| Landing low quality |

The classification performance evaluation index of random forest classification prediction model is constructed as follows.

- Classification accuracy (Accuracy): the overall classification accuracy of the test set by the random forest model

\[
Accuracy = \frac{THQ + TLQ}{THQ + FLQ + TLQ + FHQ}
\]  

- Sensitivity: the high classification accuracy of landing quality by the random forest classification model

\[
Sensitivity = \frac{THQ}{THQ + FLQ}
\]  

- Specificity: the low classification accuracy of landing quality by the random forest classification model

\[
Specificity = \frac{TLQ}{TLQ + FHQ}
\]

3. Empirical Research

This study mainly collected data of 19 flight parameters of a certain type of UAV (Unmanned Aerial Vehicle) during 47 landing flight phases. The 19 flight parameters variables are the flight distance, atmospheric altitude, vacuum speed, ground speed, roll angle, pitch angle, aileron displacement, true heading, pitch rate, yaw rate, rudder displacement, roll rate, lift speed, forward acceleration, side offset, normal acceleration, elevator displacement, lateral acceleration, engine speed.

There may be information overlap and redundancy in the 19 flight parameters. Therefore, this paper uses neighborhood rough set to flight parameters attribute reduction. Data standardization is as follows:

| Model classification prediction results. |
|-----------------------------------------|
| Record indicators | True situation | Result               |
|-------------------|----------------|----------------------|
| THQ               | Landing high quality | Landing high quality |
| FHQ               | Landing low quality    | Landing high quality |
| TLQ               | Landing low quality    | Landing low quality   |
| FLQ               | Landing high quality   | Landing low quality   |

The classification performance evaluation index of random forest classification prediction model is constructed as follows.

- Classification accuracy (Accuracy): the overall classification accuracy of the test set by the random forest model

\[
Accuracy = \frac{THQ + TLQ}{THQ + FLQ + TLQ + FHQ}
\]  

- Sensitivity: the high classification accuracy of landing quality by the random forest classification model

\[
Sensitivity = \frac{THQ}{THQ + FLQ}
\]  

- Specificity: the low classification accuracy of landing quality by the random forest classification model

\[
Specificity = \frac{TLQ}{TLQ + FHQ}
\]
Table 2. Flight data standardization.

| Flight phase | Flight height | $C_1$ | $C_2$ | ...... | $C_{19}$ |
|--------------|--------------|-------|-------|--------|---------|
| 1            | 9            | 0.7415| 0.6217| ......  | 0.7983  |
| 1            | 8.5          | 0.7330| 0.6183| ......  | 0.8067  |
| :           | :            | :     | :     | :      | :      |
| 1            | 2            | 0.1676| 0.2162| ......  | 0.9160  |
| 2            | 9            | 0.7417| 0.6588| ......  | 0.7275  |
| 2            | 8.5          | 0.7330| 0.6554| ......  | 0.7395  |
| :           | :            | :     | :     | :      | :      |
| 2            | 2            | 0.1364| 0.1926| ......  | 0.6975  |
| :           | :            | :     | :     | :      | :      |
| 47           | 2            | 0.3835| 0.3480| ......  | 0.3613  |

The standard deviation of the standardized flight parameters, the standard deviation of 19 flight parameters can be obtained as table 3.

Table 3. Standard deviation of flight parameters.

| Attribute | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ | $C_6$ | $C_7$ | $C_8$ | $C_9$ | $C_{10}$ |
|-----------|-------|-------|-------|-------|-------|-------|-------|-------|-------|----------|
| std($C_i$)| 0.149 | 0.262 | 0.153 | 0.199 | 0.218 | 0.132 | 0.104 | 0.108 | 0.248 | 0.128    |
| Attribute | $C_{11}$ | $C_{12}$ | $C_{13}$ | $C_{14}$ | $C_{15}$ | $C_{16}$ | $C_{17}$ | $C_{18}$ | $C_{19}$ | —        |
| std($C_i$)| 0.161 | 0.139 | 0.179 | 0.166 | 0.136 | 0.159 | 0.098 | 0.242 | 0.201 | —        |

The radius of the neighborhood of each parameter can be obtained by the following formula.

$$\delta(C_i) = \frac{Std(C_i)}{\lambda} \quad (5)$$

Table 4. Characteristic parameters under different $\lambda$ condition.

| $\lambda$ | Flying parameters |
|-----------|-------------------|
| 0.5       | $C_1, C_3, C_4, C_6, C_7, C_8, C_{10}, C_{11}, C_{15}, C_{16}, C_{17}, C_{19}$ |
| 0.7       | $C_3, C_7, C_8, C_{11}, C_{12}, C_{13}, C_{14}, C_{15}, C_{17}, C_{19}$ |
| 0.9       | $C_3, C_4, C_7, C_8, C_{11}, C_{14}, C_{15}, C_{19}$ |
| 1.1       | $C_3, C_{11}, C_{14}, C_{15}, C_{16}, C_{17}, C_{19}$ |
| 1.3       | $C_3, C_7, C_{11}, C_{17}, C_{19}$ |
Among them, \( Std(C_i) \) represents the standard deviation of the parameter \( C_i \) of the flight attribute, and the difference of \( \lambda \) directly affects the size of the radius of the domain, thus affecting the number of reduced attributes.

The value of \( \lambda \) and the lower limit of importance degree were taken as 0.01 to reduce the properties of the flight attributes shown in table 4.

At the same time, this paper adopts factor analysis for attribute reduction as a comparative analysis. The results of the factor analysis of the flight parameters are shown in table 5.

### Table 5. Flight parameters analysis.

| Flying parameters | Total | Original feature value | Extract the square load |
|-------------------|-------|------------------------|------------------------|
|                   |       | Variance ratio (%)     | Cumulative ratio (%)   |
|                   |       | All                    | Variance ratio (%)     | Cumulative ratio (%) |
| 1                 | 4.163 | 21.913                 | 4.163                  | 21.913               |
| 2                 | 2.769 | 14.572                 | 2.769                  | 14.572               |
| 3                 | 1.732 | 9.117                  | 1.732                  | 9.117                |
| 4                 | 1.659 | 8.734                  | 1.659                  | 8.734                |
| 5                 | 1.424 | 7.493                  | 1.424                  | 7.493                |
| 6                 | 1.107 | 5.826                  | 1.107                  | 5.826                |
|                   |       | 67.655                 | 67.655                 |

Select the dimensionality reduction factor with an eigenvalue greater than 1 as the flight parameter. According to table 4, we can get the original flight parameter information of 67.655% with the six flight parameters retained by 19 flight parameters after dimensionality reduction. The six flight parameters that are reserved are defined as \( f_1, f_2, f_3, f_4, f_5, f_6 \).

This paper model input variables can be roughly divided into three categories. The first category is the original flight variable, which is 19 flight variables; the second category is the flight parameter variable after the neighborhood rough set reduction under different \( \lambda \)-valued conditions, 12, 10, 8, 7 and 5 variables respectively; the third category is the six parameters obtained after factor analysis.

The forecast precision of random forest model constructed under each input variable is shown in table 6.

### Table 6. Comparison of Prediction Results of Different Input Variables.

| Enter the number of variables | OOB Error rate (%) | Accuracy (%) | Sensitivity (%) | Specificity (%) |
|------------------------------|-------------------|--------------|----------------|-----------------|
| 19                           | 23.20             | 76.80        | 80.40          | 69.60           |
| 12                           | 20.53             | 79.47        | 81.60          | 75.20           |
| 10                           | 19.87             | 80.13        | 82.40          | 75.60           |
| 8                            | 16.67             | 83.33        | 84.80          | 80.40           |
| 7                            | 13.33             | 86.67        | 88.20          | 83.60           |
| 5                            | **8.27**          | **91.73**    | **92.80**      | **89.60**       |
| 6                            | 16.37             | 83.60        | 86.80          | 77.20           |
4. Conclusions
The random forest classification prediction model constructed by the input variables of the reduced dimension of the neighborhood rough set and factor analysis is more effective than the random forest classification prediction model constructed from the original variables, indicating that there is redundant information among the original variable data which affects the prediction accuracy of the model.

The effect of stochastic forest classification prediction model constructed by factor analysis dimensionality reduction is better than that of stochastic forest classification prediction model constructed by attribute reduction of neighborhood rough set in $\lambda = 0.5, 0.7, 0.9$ condition, but it is better than the attribute reduction of neighborhood rough set in $\lambda = 1.1, 1.3$ condition. The random forest classification prediction model has poor effect, which shows that $\lambda$ is an important factor affecting the attribute reduction effect of neighborhood rough set. However, the attribute reduction algorithm of neighborhood rough set in this paper can deal with the flight parameter data well. The random forest classification prediction model constructed by five input variables obtained by reducing the attributes of neighborhood rough set under the condition of height 8m, $\lambda = 1.3$, has the best effect. The OOB error rate of the training set is 8.27%, while the overall accuracy of the model is Accuracy 91.73%. Sensitivity of the model was 92.80%, specificity of the model was 89.60%. Which can well predict the landing landing quality is high or low, and the prediction accuracy is high, with high credibility, to provide guidance for the landing of the aircraft.

Acknowledgements
The authors acknowledge the National Natural Science Foundation of China (Grant: 71501007 & 71672006). The study is also sponsored by the Technical Research Foundation and the Graduate Student Education & Development Foundation of Beihang University. There is no conflict of interest to declare.

References
[1] Cross E J, Sartor P and Worden K 2013 J. Structural Health Monitoring 18 1452-1459
[2] Lan C E, Wu K and Yu J 2012 J. Chinese Journal of Aeronautics 23 511-517
[3] Wang L, Sun R and Wu C X 2014 J. Chinese Journal of Safety Science 18 1452-1459
[4] Pawlak Z 1982 J. International Journal of Information and Computer Science 11 129-141
[5] Wang X and Yan X Y 2017 J. Journal of Aeronautics 38 71-76
[6] Khan M I 2018 J. Journal of Cleaner Production 172 1035-1052
[7] Yazdani M and Payam A F 2016 J. Materials & Design 65 328-334
[8] Xu F, Liu J and Lin S 2017 J. Journal of Cleaner Production 148 254-267
[9] Gul M, Celik E and Aydin N 2016 J. Applied Soft Computing 46 60-89
[10] Hua H H, Yu D R and Xie Z X 2008 J. Numerical Journal of Software 19 640-649
[11] Breiman L 2001 J. Machine Learning 45 5-32
[12] Tetschke F, Schneider U and Schleussner E 2016 J. Computers in Biology and Medicine 70 157-162
[13] Zhao T, Wang L T and Zhang Y 2016 J. Chinese Society for Electrical Engineering 36 604-614
[14] Wu Y Y, He J H and Zhang P 2015 J. Automation of Electric Power Systems 39 50-55
[15] Ding J M, Liu G Q and Li H 2015 J. Pattern Recognition and Artificial Intelligence 28 1041-1049
[16] Zhu Z J, Zhang H W and Chen H 2015 J. Hydrogeology and Engineering Geology 42 108-113