A mechanical wearing fault diagnosis method of aero-engine based on RSS-ERT

Mingyang Tang*, Yafeng Wu
School of Power and Energy, Northwestern Polytechnical University, Xi'an 710129, China

ABSTRACT

A fault diagnosis method based on RSS-ERT was proposed to deal with the mechanical wearing fault of aero-engine through combining with oil analysis. After pre-processing the element concentration data of aero-engine oil sample analysis, in order to train more models and have more randomness, Return Sampling Strategy (RSS) was adopted in the build of Extreme Random Tree (ERT) model. The coefficient of determination (R^2) was used to evaluate performance of the model. This model and benchmark model were used to forecast engine health parameters. The result showed that the fault diagnosis method based on RSS-ERT was accurate and superior.

Keywords: Aero-engine, extreme random tree, oil analysis, bagging, fault diagnosis

1. Introduction

Aero engine is the core component of aircraft. It has extremely complex system and structure[1]. It is prone to mechanical wear failure in the face of harsh operating environment and long working time. The lubricating oil system can effectively reduce the friction between engine parts. If the metal content in the lubricating oil component increases significantly, it generally indicates that some parts of the engine have serious mechanical wear[2]. Therefore, the wear of the engine system is usually mapped through oil analysis. Machine learning, as the most popular research topic in the field of artificial intelligence[3], can mine the potential connections within a large amount of data through various algorithms[4], which has great advantages in building prediction models.

In this paper, an ensemble algorithm combined with oil-liquid analysis is selected to design a fault diagnosis method based on return sampling strategy extreme random tree (RSS-ERT) to predict engine wear status. The data set is divided into training set and test set, and then a fault diagnosis model based on extreme random tree algorithm is established. Compared with random forest algorithm, this algorithm avoids the over-fitting of training model and the optimization of complex parameters, and improves the accuracy. This method not only has stronger randomness of the training model, but also improves the generalization performance. The coefficient of determination (R^2) was used to evaluate performance of the model, and fault diagnosis can be realized by inputting the oil sample data of an aero engine into the trained model. The results show the effectiveness and superiority of the proposed algorithm.

2. Decision Tree

2.1 Basic model

Decision tree is a widely used inductive reasoning algorithm. In classification problems, decision tree algorithm classifies samples based on features to form a 'tree' containing a series of if-then rules. Mathematically, this tree can be interpreted as a conditional probability distribution defined in feature space and class space[5].

In Figure 1, many nodes and directed edges are combined to form a "tree". Nodes can be roughly divided into two categories: internal nodes (in) and leaf nodes (LN). The IN indicates a feature, and the LN represents a classification mark. If the current node is an internal node, it moves to a child node of the current node according to the value of the feature corresponding to the sample, and the child node corresponds to a value of the feature. In this way, the recursion will finish in the leaf node, and the classification mark represented by the leaf node is returned. That's a whole classification process.

*tmy2021100232@mail.nwpu.edu.cn

International Conference on Computer Application and Information Security (ICCAIS 2021),
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12260, 122600Q · © The Authors.
Published under a Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.2637623

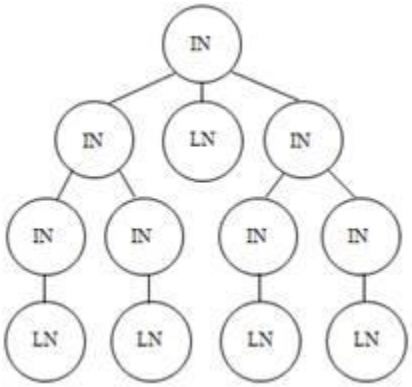

Figure 1. Decision tree.

2.2 Generating regression decision tree

The key to generating the decision tree is the selection of the best segmentation features. In this paper, the mean square error (MSE) is selected to generate the regression tree. We set the input variable to X. Y is the mapping of X as the output variable. The training data set is $K = \{(x_1, y_1), (x_2, y_2) \ldots (x_m, y_m)\}$. Then a decision tree can be built recursively and the specific steps are as follows:

(1) The optimal partition variable j and the partition point s are selected, and two regions $E_1(j,s) = \{x \mid x^{(j)} \leqslant s\}, E_2(j,s) = \{x \mid x^{(j)} > s\}$ are divided. You need solve the equation (1).

$$\min_{j,s}\left[\min_{c_1}\sum_{x_i \in \mathrm{E}(j,s)} (y_i - b_1)^2 + \min_{c_2}\sum_{x_i \in E_2(j,s)} (y_i - b_2)^2\right] \quad (1)$$

$$b_1 = \frac{1}{n}\sum_{x_i \in E_1(j,s)} y_i, \quad b_2 = \frac{1}{m-n}\sum_{x_i \in E_2(j,s)} y_i$$

n is count of x in $E_1(j,s)$. The pair (j, s) is selected that minimizes equation (1).

(2) The area is divided with the selected pair (j, s)， and the corresponding output value is obtained by minimizing MSE:

$$\hat{b_m} = \frac{1}{N_m}\sum_{x_i \in E_m(j,s)} y_i, x \in E_m, m = 1,2 \quad (2)$$

(3) Calling Steps (1) and (2) for the two sub areas. We set a termination condition. When the termination condition is met, Step (3) stops

(4) Dividing the input space into M regions E_1, E_2... E_M to generate a decision tree:

$$f(x) = \sum_{m=1}^{M} \hat{b_m} L(x \in E_m) \quad (3)$$

L is an indicator function: $L = \begin{cases} 1 & if\ (x \in E_m) \\ 0 & if\ (x \notin E_m) \end{cases}$.

3. RSS-ERT

Extreme random tree is not only a machine learning algorithm based on Bagging, but also a variant of random forest algorithm. It uses the whole data set as samples each time, and uses CART decision tree[6] as the basic weak learner model. Each decision tree is independent of each other. In the final model combination process, the final result is the arithmetic mean of all decision tree models for the regression problem[7].

RSS can make it easier to train more models with more random. It may lead to some samples not being taken. According to statistics, about 37% of the samples are not taken. This part of the samples is called out of bag samples (OOB)[8]. Using OOB as the test set to verify the model is more convincing, and the results of the decision tree model can be weighted to increase its accuracy.

The algorithm flow of RSS-ERT is shown in Figure 2. Overall sample set is $D = \{(X_1, y_1), (X_2, y_2) \ldots (X_N, y_N)\}$. X_i is $1 \times Q$ dimension row vector, y_α is x_α the true output value of the corresponding sample, $\alpha = 1,2 \ldots N$, N is the number of sample groups, and the algorithm steps are as follows:

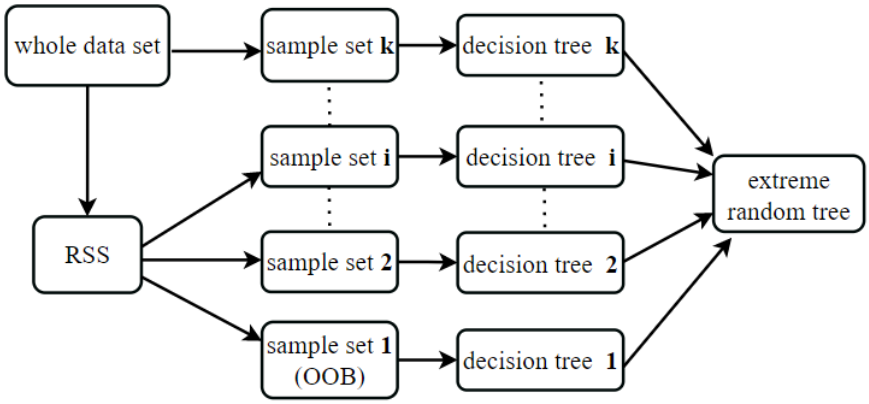

Figure 2. RSS-ERT algorithm.

(1) Assuming that the total number of training sample sets is K, it is divided into two parts. One part samples sample set D by RSS, and the number of sample sets is *I*; the other part directly uses the whole sample set, and the number of sample sets is *K-I*.

(2) In the classification model of extreme random tree, the base classifier uses the training sample sets generated in Step (1) for training. Compared with the traditional extreme random tree algorithm, this process reduces the amount of training, and provides a guarantee for the independence of the base classifiers in the next step to get better prediction results.

(3) Number of features is *W*. The base classifier is built by the algorithm in Section 2.2. To amplify randomness, when splitting each node, not only m features from W features should be selected randomly, but also the optimal attributes should be selected, and pruning is not allowed throughout the process. The data subset generated by splitting is iterated Step (3). When the "tree" is built, Step (3) stops.

(4) Repeat Step (2) and Step (3) for *K* times to generate *K* decision trees.

(5) Prediction results are generated by using OOB samples for the generated extreme random tree, and they are weighted to increase accuracy.

4. EXPERIMENT AND RESULT ANALYSIS

4.1 Experimental data

In this paper, the model is trained based on the element concentration data of a military aero engine oil sample analysis, and the engine wear state is predicted. There are 212 samples in the data set, including the lubricating oil state of the engine under normal state and wear state. The concentration values of Fe, Al, Cu, Cr, Ag, Ti and Mg are attributes, and *F* as a label is the fault type of the engine. "1" represents normal state, "2" represents inter shaft bearing wear and "3" represents inter shaft bearing wear with cage fracture. Some data are shown in Table 1.

Table 1. Element concentration data of a military aero engine oil sample (part).

Fe	Al	Cu	Cr	Ag	Ti	Mg	F
4.8	0	1.5	0.2	0.1	1	6.1	1
16	0.5	2.4	1.4	0.5	1.1	7.2	2
1.6	0	0.7	0	0	0.6	3.3	1
1.6	0	0.8	0	0.1	0.8	3.4	1

In this paper, 30% of the whole data set is used as the test set and 70% as the training set. Due to the large difference in the number of samples under each label, the class imbalance problem is caused. Therefore, the hierarchical sampling technology is adopted. The distribution of data samples is shown in Table 2.

Table 2. Sample distribution of experimental data.

F	Number of training sets	Number of test sets
1	144	61
2	3	2
3	1	1

4.2. Experimental results and analysis

In this paper, the RSS-ERT model is used to estimate the influence (importance) of the concentration of Fe, Al, Cu, Cr, Ag, Ti and Mg on the accuracy of predicting engine health parameters. After normalizing the importance of these seven elements, they are arranged in descending order according to the importance, as shown in Figure 3.

The concentration values of the above seven elements in the lubricating oil are 1.0, 0.32, 0.43, 0.38, 0.28, 0.03 and 0.06. It can be seen from Figure 3 that the concentration of Fe in the lubricating oil has a great impact on the engine health. The concentration values of Cr, Ag and Al have little difference, and the degree of impact is not high. The concentration value of Ti is the least important.

The evaluation standard of the model is the determination coefficient R^2[29] and the mathematical expression of R^2 is as follows.

$$R^2(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2} \tag{4}$$

y is the actual value; $\hat{y}$ is predicted value; $\overline{y}$ indicates the average of the predicted values of sample points, and N is the amount of samples. The values of R^2 closer to 1 indicate better prediction. This paper selects the typical linear regression model[10] and random forest model[11] as the reference, and obtains the R^2 of the model through the 50% cross validation of the data set. The values are shown in Figure 4.

In Figure 4, the R^2 of the linear regression model is the lowest, at 0.244, which also reflects that the linear model prediction is not suitable for the engine oil fault diagnosis in this experiment, and the integrated algorithm has more advantages. In the integrated algorithm, the random forest has been significantly improved to 0.724. The RSS-ERT model proposed in this paper is the highest (0.972), which is very close to 1, and the prediction effect is the best.

In this model, the impact of the number of decision trees on the prediction effect is shown in Figure 5. In the experiment, the prime number of decision trees is 50, and then increased by 10 every time until 500. It can be seen from Figure 5 that the impact of the increase of the number of decision trees on the prediction effect fluctuates in the initial stage, and gradually stabilizes in the later stage, which the R^2 is 0.969 and the lowest value is 0.954, so it reflects the superiority of this model.

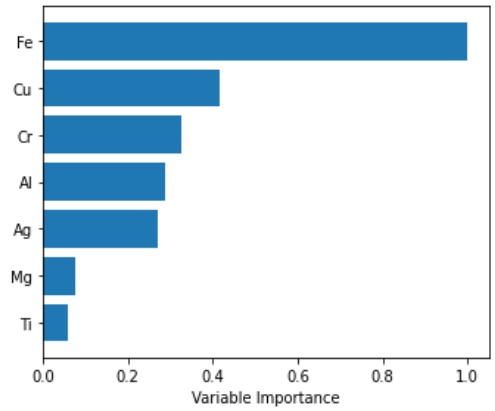

Figure 3. Importance ranking of seven elements.

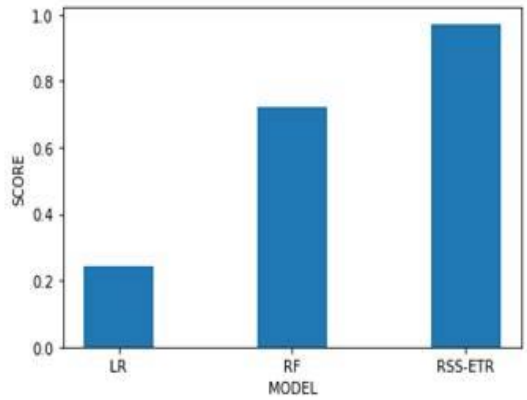

Figure 4. Evaluation indexes of different models.

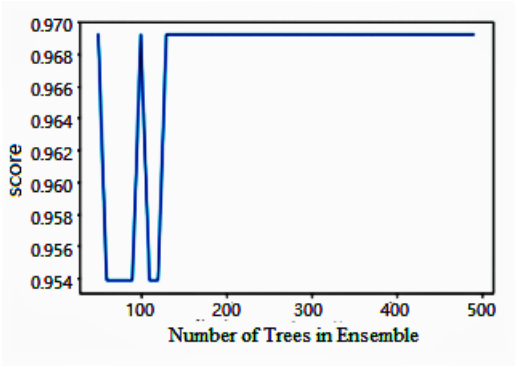

Figure 5. Overall performance of RSS-ERT model.

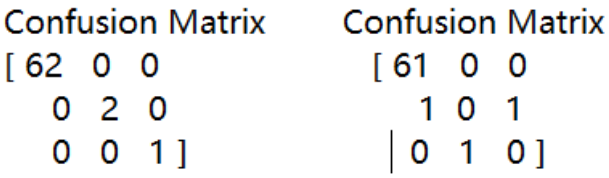

Figure 6. Confusion matrix predicted by RSS-ERT model and random forest model.

In order to more intuitively see whether the model predicts the correct results, the confusion matrixes[12] of RSS-ERT model and random forest model are obtained. In Figure 6, it indicates that the predicted value of RSS-ERT model is consistent with the measured data, and the prediction results are accurate. The predicted value of the random forest model is less consistent with the measured data than the RSS-ERT model. One of them is not predicted in the state "1", and the prediction of the state "2" and "3".

According to Figures 4-6, the prediction accuracy of the model generated by this algorithm is higher.

5. CONCLUSIONS

For the faults caused by mechanical wear of aero engine, combining with oil analysis, a fault diagnosis method based on RSS-ERT is proposed in this paper. Compared with the traditional ensemble learning algorithm, the classification model reduces the amount of training, and provides a guarantee for the independence between the base classifiers. This method has stronger randomness of the training model, improves the generalization performance, and has a wide application prospect. It is applied to the prediction of wear fault of a military aero engine, the results show that the RSS-ERT algorithm has better performance and higher prediction accuracy.

REFERENCES

[1] Cui, J. G., Li, Y., Cui, X., Wang, J. L., Jiang, L. Y. and Yu, M. Y., "An improved fault diagnosis method for lubricating oil system of DBN aero engine," Journal of Shenyang University of Aeronautics and Astronautics**,** 37(06), 49-54 (2020). (in Chinese)

[2] Zhang, C. Y., “Research on diagnosis of aero engine mechanical wear fault,” Equipment Management and Maintenance, (24), 147-148 (2020).
[3] Wu, J. D., [Research on Machine Learning Method for Aero Engine Gas Path Fault Diagnosis and Prediction], Nanjing University of Aeronautics and Astronautics, Nanjing, Master’s Thesis, (2019). (in Chinese)
[4] Robert, C., “Machine learning: A probabilistic perspective,” Chance, (2), (2014).
[5] Cheng, Y., Li, Y. X. and Li, F., “Soil moisture inversion in lightning river basin based on extreme random tree,” Journal of Remote Sensing, 25(04), 941-951 (2021).
[6] Wang, Y., Chen, D. Y. and Tang, Y. X., “Stock forecasting based on cart decision tree and boosting method,” Journal of Harbin University of Technology, 24(06), 98-103 (2019). (in Chinese)
[7] Breiman, L., “Random forests,” Machine Learning, 45(1), 5-32 (2001).
[8] Zhang, C. L., Yang, G. J., Li, H. L., Tang, F. Q., Liu, C. and Zhang, L. Y., “Remote sensing inversion of winter wheat leaf area index based on random forest algorithm,” China Agricultural Science, 51(05), 855-867 (2018).
[9] Zhao, S. H., “Analysis and evaluation of influencing factors on goodness of fit R^2,” Journal of Northeast University of Finance and Economics, (03), 56-58 (2003).
[10] Chen, X., “Analysis on influencing factors of air quality index based on linear regression model—Taking Dazu District of Chongqing as an example,” Environmental Impact Assessment, 43(05), 79-82 (2021).
[11] Zhu, P. J., Luo, N. X. and Zhao, Q. S., “Prediction of maximum water increase of typhoon storm surge based on stochastic forest model,” Surveying and Mapping Bulletin, (12), 71-74 + 82 (2021).
[12] Gao, Y. and Peng, W., “Research on Application of classification prediction based on keras,” Journal of Shanxi Datong University (Natural Science Edition), 35(05), 26-30 (2019).