Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Development of fault diagnosis platform for liquid rocket engine based on experimental data

To cite this article: Shoujin Chang *et al* 2021 *J. Phys.: Conf. Ser.* **2093** 012022

Development of fault diagnosis platform for liquid rocket engine based on experimental data

Shoujin Chang[1], Yong He[2], Haitao Hu[1*], Jiaxin Chai[2], Minghe Hu[1], Chi Zhang[1]

[1] Institute of Refrigeration and Cryogenics, Shanghai Jiao Tong University, Shanghai 200240, China

[2] Beijing Aerospace Automatic Control Institute, Beijing 100854, China;

[*]Corresponding author's e-mail: huhaitao2001@sjtu.edu.cn

Abstract. Aiming at the defects of difficult to establish fault diagnosis models caused by the small scale of liquid rocket engine test data, imbalanced categories and high fault coupling, the fault diagnosis platform with 10 fault diagnosis models are developed for fault diagnosis, covering K-nearest neighbor (KNN) model, optimized KNN model, logistic regression (LR) model, optimized LR model, support vector machine model, K-means model, decision tree model, neural network model, random forest model and light-GBM model. The prediction accuracies of these models are validated based on the experimental data. Among these models, the light-GBM model provide the best prediction accuracies, and 5 models have the prediction accuracy larger than 98%.

1. Introduction

Liquid rocket engine is the core part of liquid rocket. Once it breaks down, it will cause a lot of financial losses. Therefore, the real-time diagnosis on the state of the engine and finding the abnormal signs of the engine timely and accurately is of great significance for equipment maintenance personnel to improve the operation safety of the engine[1].

At present, the fault diagnosis model of liquid rocket engine is mainly divided into model-based method, data-based method and artificial intelligence method[2]. These methods have been widely used in the fault diagnosis of liquid rocket engine, and have played a great role and is of great value. Han et al. [3] applied the decision tree to extract the fault features and carried out the fault detection and diagnosis in the steady-state section; Huang et al. [4] developed a BP algorithm for real-time fault detection based on neural network technology for a large liquid rocket engine; He et al. [5] used support vector machine to detect and diagnose the fault of liquid rocket engine, and verified the correctness and reliability. However, the models may not be suitable for the liquid rocket engine based on small scale of liquid rocket engine test data, and there is no fault diagnosis platform for liquid rocket engine.

In the present study, a fault diagnosis platform with different models will be developed and be validated based on the experimental data.

2. FAULT DIAGNOSIS DATA AND ALGORITHM EVALUATION MODEL

2.1. Fault diagnosis data

The data used in this paper were provided by Beijing Aerospace Automatic Control Institute. There are 264 sets of sample data in total, and each set contains 10 monitoring parameters ($x1$-$x10$). The normalized data are shown in Table 1.

Table 1. Normalized data.

$x1$	$x2$	……	$x9$	$x10$	$y1$	$y2$
0.8805	0.9818	……	-0.2926	0.6672	0	0
0.8589	0.9170	……	-0.3724	0.6619	1	11
0.8354	0.8479	……	-0.4502	0.6566	1	11
0.8906	0.7744	……	-0.5267	0.6512	1	11
0.7812	0.6960	……	-0.6014	0.6449	1	11
……	……	……	…….	……	……	……
0.5766	0.8432	……	-0.5457	0.5375	4	42
0.5502	0.8269	……	-0.5641	0.5279	4	42
0.8362	0.6802	……	0.8792	0.9819	4	43
0.8197	0.6611	……	0.9380	0.9904	4	43
0.8018	0.6410	……	1.0000	1.0000	4	43

2.2. Precision, recall and F1_score

Precision and recall are two metrics widely used in the fields of information retrieval and statistical classification to evaluate the quality of results. In the field of fault diagnosis, the comparison samples are usually normal samples and fault samples, so the prediction situation can be divided into the following four types:

1) TP: predict normal samples as normal samples.
2) FN: predict normal samples as fault samples.
3) FP: predict failure samples as normal samples.
4) TN: predict failure samples as failure samples.

Among them, the precision is the ratio of the number of retrieved related samples to the total number of retrieved samples, which measures the accuracy of the diagnostic model. The recall rate refers to the ratio of the number of retrieved related samples to the number of all related samples in the sample set, which measures the recall rate of the diagnostic model. According to the above definition, the precision rate, the recall rate and F1_score are defined as follows:

$$precision = \frac{TP}{TP+FP} \tag{1}$$

$$recall = \frac{TP}{TP+FN} \tag{2}$$

$$F1 = \frac{2TP}{2TP+FN+FP} \tag{3}$$

F1_score is the harmonic mean of precision and recall. When the precision and recall are both high, F1_score will also be at a higher level.

2.3. Receiver operating characteristic curve

The ROC (receiver operating characteristic) curve more intuitively reflects the characteristics of accuracy and recall. Its horizontal axis represents the false positive rate (FPR), representing the proportion of the actual normal samples in all normal samples predicted by the diagnostic model. The vertical axis represents the true positive rate (TPR), which means the proportion of the actual faulty samples in all faulty samples predicted by the diagnostic model. The calculation formulas are:

$$\text{TPR} = \frac{TP}{TP+FN} \tag{4}$$

$$\text{FPR} = \frac{FP}{FP+TN} \tag{5}$$

The ROC curve actually describes the process in which the performance of the classifier changes with the change of the classifier threshold. For the ROC curve, an important feature is its area. The area is 0.5 for random classification, and the recognition ability is 0. The closer the area is to 1, the stronger the recognition ability is. The area is equal to 1 for complete recognition. The ideal target is TPR=1, FPR =0. The more the angle of the curve deviates from the 45-degree diagonal, the better. The typical ROC curve diagram is shown in Figure 1.

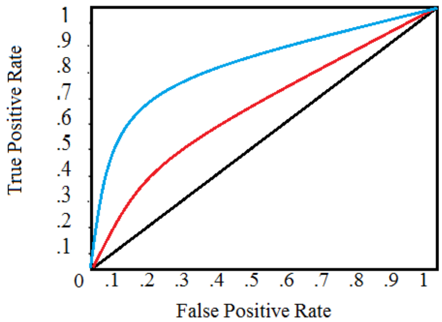

Figure 1. Receiver operating characteristic curve.

3. FAULT DIAGNOSIS MODELS AND VALIDATIONS

3.1. Fault diagnosis platform and models

In order to facilitate user operation, call model, training model and data visualization operation, a fault diagnosis platform is developed for data import, model selection, and data analysis. In the fault diagnosis platform, 10 fault diagnosis models are developed, and are validated based on the normalized data shown in Table 1.

3.1.1. KNN (K- Nearest Neighbor) model and optimized KNN model

KNN (K- Nearest Neighbor) has no explicit training process. During training, the distance between the test sample and all training samples is calculated. Given a test sample x, its nearest k training examples form the set $N_k(x)$, and the classification loss function is 0-1 loss. The first 70% of the data is taken as the training set, and the remaining 30% is taken as the test set. The change curve of the algorithm accuracy rate is shown in Figure 2. As k gradually increases, the overall accuracy rate initially increases and then decreases, representing the higher values of 88.7% at k of 5~8.

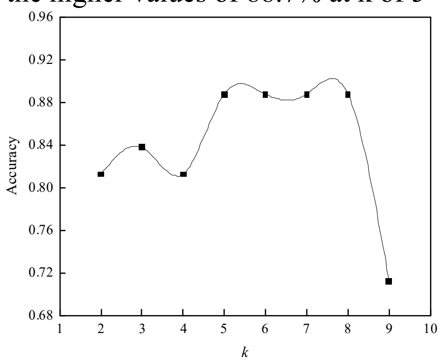

Figure 2. Change curve of accuracy rate under different values of k.

For improving the prediction accuracy, an optimization KNN model is proposed by setting two additional hyperparameters of the distance weight and the distance p value; a distance weight is set, so that the classification standard is no longer only the number of neighbors; Minkowski distance is used in the optimized model, as shown in Eq. (6). After training, the final accuracy of optimized KNN model is 98.40%.

$$(\sum_{i=1}^{n} | X_i^{(a)} - X_i^{(b)} |^p)^{\frac{1}{p}} \tag{6}$$

where, X_i represents the sample independent variable, and a and b represent different samples.

3.1.2. LR (Logistic Regression) model and optimized LR model

LR adds a Sigmoid function mapping on the basis of linear regression[6]. The training method is mainly to combine any two of the n data classification results, and then train and predict them separately. Finally, among all the prediction categories, the one with the highest win number is the classification result. The parameter matrix of the LR multi-classification model obtained through training is 10×5 dimensions, and the specific values are shown in Table 2. Through the parameter matrix obtained above, the calculation accuracy of LR for the multi-classification problem is 80.60%. The problem of the original LR model is caused by the unbalanced data sets.

Table 2. LR multi-class model parameter matrix.

0.21	0.11	-0.07	0.04	0.08	-0.01	0.08	0.05	0.23	0.07
-0.03	-2.25	-0.24	1.48	1.72	-0.79	-0.51	-0.62	-0.10	-0.31
0.25	1.22	0.21	-0.47	-1.10	-0.94	2.54	1.34	0.48	0.82
0.37	0.75	0.39	-0.08	-0.72	-1.13	-0.71	-0.85	-0.04	-1.80
-0.79	0.17	-0.29	-0.96	0.02	2.87	-1.40	0.08	-0.57	1.21

An optimized LR model is proposed by directly considering the fault category, because it intersects with the fault category and the fault is small. After training, the final accuracy of optimized LR model is 98.40%.

3.1.3. SVM (Support Vector Machine) model

SVM (Support Vector Machine) is a binary classification algorithm, but regression can also be done[5]. One VS One training method is used to transform it into a multi-classification problem. during the training process, the penalty parameter of SVC is selected as 1, the kernel function is selected as Gaussian kernel function. SVM generates multiple support vectors during the training process, and the number of support vectors in each category is shown in Table 3. More support vectors are required for the judgment of category 2 and category 4, which indicates the bottleneck of SVM for judging these two types of faults. Through training, the calculation accuracy of SVM for this multi-classification problem is 79.10%.

Table 3. Number of support vectors for different types of faults.

type	0	1	2	3	4
Number of support vectors	1	16	67	12	78

3.1.4. K-Means model

The core idea of the algorithm for K-Means model is to classify the clustering center[7]. A fixed number of cluster centers are selected as the fault center, and the new data are classified into the specific fault according to smaller distance. According to the calculation, the cluster center matrix is:

$$\begin{bmatrix} 0.40336 & 0.303695 & -0.17349 & -0.01519 & 0.161005 & 0.12831 & 0.01826 & 0.01991 & 0.502135 & 0.450505 \\ 0.74474375 & 0.88729125 & -0.03172875 & -0.245375 & -0.1996925 & 0.3249075 & 0.30089875 & 0.27674375 & 0.8034775 & 0.70488125 \\ 0.14803125 & 0.88685625 & -0.0467875 & -0.59271875 & -0.53219375 & -0.41720625 & -0.4809375 & -0.4864125 & 0.0588375 & -0.3017 \\ 0.43747917 & 0.81026042 & -0.47212708 & -0.56117604 & 0.14018021 & 0.30015625 & 0.20156667 & 0.20486042 & 0.56799375 & 0.50889479 \end{bmatrix}$$

The accuracy of K-Means on the training set was 27%. K-Means considers the importance of all features to be the same in the calculation process, and each feature has the same weight in the distance. So the final distance will be interfered by low-importance features, which leads to the low accuracy.

3.1.5. Decision tree model
Decision tree model is based on the probability of occurrence of various situations. By forming a decision tree, it obtains the probability that the expected value of the net present value is greater than or equal to zero[5]. The algorithm adopts a tree structure and uses layered inference to achieve the final classification. It is a supervised learning algorithm based on if-then-else rules. These rules of the decision tree are obtained through training instead of manual formulation.

All data are randomly scrambled and then screened to ensure the data integrity of all kinds of data in the training model. According to the training, the accuracy rate of the final decision tree model is 98.4%.

3.1.6. Neural Network model
The model uses a two-layer neural network, including a single hidden layer, an input layer, and an output layer. The number of nodes in the input layer is 10, the nodes in the hidden layer are 8, and the nodes in the output layer are 5. The activation function after the hidden layer is Relu function, a dropout layer with p=0.2 is added, and the optimization algorithm is Adam algorithm. The neural network structure after training is shown in Figure 3. The average accuracy of the model is 70.36%. Neural network is not suitable for small-capacity data.

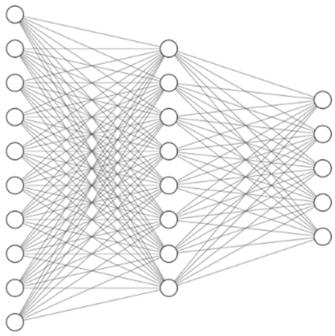

Figure 3. Schematic diagram of neural network structure.

3.1.7. Random forest model
Because the weak classifier of the random forest is a decision tree, the random forest model can inherit the good results of the decision tree model for solving this problem, and improve the generalization of the model through randomness[8]. CART (Classification and Regression Trees) algorithm is used to construct a decision tree in a random forest, and the number of decision trees in the forest is set as 100; the maximum depth of each tree is 6, and the number of training iterations is set as 1000. After training, the final accuracy of random forest model is 98.14%.

3.1.8. LightGBM model
LightGBM is an efficient implementation of GBDT (Gradient Boosting Decision Tree)[9]. The Random Forest algorithm is used as the basic model of light-GBM, the number of decision trees is set to 100, the maximum depth of each tree is After model training, the accuracy of the model for fault classification can reach up to 100%, which is improved on the basis of the decision tree model.

3.2. Model validation results
The accuracies of the developed models validated based on the experimental data are summarized and listed in Table 4. Among these models, the light-GBM model provide the best prediction accuracies, and 5 models have the prediction accuracy larger than 98%.

Table 4. The accuracy of the models in the fault diagnosis platform.

Model	Prediction accuracy
KNN	88.7%
LR	80.6%
SVM	79.1%
K-Means	27%
Decision tree	98.4%
neural network	70.36%
optimized KNN	98.4%
optimized LR	98.4%
random forest	98.14%
lightGBM	100%

4. CONCLUSIONS

Based on the experimental data for liquid rocket engine, a fault diagnosis platform is developed to import data, train models and validate models; in the platform, 10 fault diagnosis models are developed, including KNN, LR, SVM, K-Means, decision tree, neural network, random forest, and light-GBM models; the models are validated based on the experimental data. The diagnostic accuracy of light-GBM can reach 100%, and 5 models have the prediction accuracy larger than 98%. The research results and platform are useful for the fault diagnosis and health management of liquid rocket engines.

Acknowledgments

This study is supported by the National Defense Basic Scientific Research Project (No. JCKY2018203B022), Shanghai Science and Technology Innovation Action Program (No. 19142203000) and State Key Laboratory of Compressor Technology (No. SKLYSJ201904).

References

[1] Cha, J., Ko, S., Park, S.Y., et al. (2019) Fault detection and diagnosis algorithms for transient state of an open-cycle liquid rocket engine using nonlinear Kalman filter methods. Acta astronautica, 163: 147-156.

[2] Wu, J. (2005) Liquid-propellant rocket engines health-monitoring—a survey. Acta Astronautica, 56(3): 347-356.

[3] Hu, X., Han, Q., Li, J. (2007) Fault Diagnosis of Liquid Rocket Engine by Dynamic Time Warping Combined with Decision Tree Method. JOURNAL-NATIONAL UNIVERSITY OF DEFENSE TECHNOLOGY, 29(4): 1-5.

[4] Yu, H., Wang, T. (2021) A Method for Real-Time Fault Detection of Liquid Rocket Engine Based on Adaptive Genetic Algorithm Optimizing Back Propagation Neural Network. Sensors, 21(15), 5026.

[5] He, H., Hu, X., Jiang, Z., Liu, W.. (2008) SVM implemented in fault diagnosis of liquid rocket engine. JOURNAL OF ROCKET PROPULSION, 3: 7-12. (In Chinese)

[6] Jing, Y., Su, H., Wang, S., Gui, W., et al. (2021) Fault Diagnosis of Electric Impact Drills Based on Time-Varying Loudness and Logistic Regression. Shock and Vibration, 2021: 1-11.

[7] Ali, A.A., Hameed, E.M., Shareef, S.M., et al. (2020) Lion Optimized K-Means Support Vector Machine for Clustering Problems in Cloud Internet of Things Environment. Solid State Technology, 63(6): 4620-4631.

[8] Cerrada, M., Zurita, G., Cabrera, D., et al. (2016) Fault diagnosis in spur gears based on genetic algorithm and random forest. Mechanical Systems and Signal Processing, 70: 87-103.

[9] Ke, G., Meng, Q., Finley, T., et al. (2017) Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30: 3146-3154.