Propulsion life prediction based on support vector machine

Ran Tang*, Gang Fang, Guangjia Liu and Heng Wang
China Satellite Maritime Tracking And Control Department
Jiangyin, 214413, China
1009127870@qq.com

Abstract. The advanced technology of the propulsion device is a national key research project, and PHM prediction and health management are among the core technologies. Our prediction of the remaining effective life (RUL) of the propulsion unit is a key part of the PHM technology. The data-driven method can effectively predict RUL. For example, the artificial intelligence method is widely used because it has certain advantages. Therefore, this paper uses machine learning algorithms to carry out RUL prediction research of propulsion devices. In addition, on the basis of support vector machine (SVM) regression theory, we combined the method of SBI (similarity-based interpolation), using Matlab program to build a prediction model. Both the practice samples and test samples of the model come from the engine simulation data, CMAPSS, provided by NASA. Finally, we use test samples to test the prediction model. The results show that the constructed model has good accuracy and robustness, and it can better predict the RUL of the propulsion unit.

1. Introduction
Failure of the propulsion device is a very serious matter. Once a failure occurs, it is difficult to estimate the damage caused to the property of the country and the people. According to statistics, 30% of ship failures are caused by propulsion units, and there are about 40% of engine failures in major accidents. The maintenance cost of the propulsion unit is extremely high, and its maintenance cost accounts for 40% of the entire ship maintenance. Replacing a new propulsion unit is as much as ten million dollars. Predicting its life expectancy can play a pivotal role in the propulsion device to reduce ship operating costs.

Scholars around the world have made extensive research on the prediction of the life of propulsion devices. Liao [1] et al. proposed a classification method of hybrid method, which is divided into 5 categories according to the combination of the above three prediction methods. H1 is based on the combination of empirical model and data drive. H2 is based on the combination of empirical model and physical model. H3 is a combination of multiple data drives. H4 is based on a combination of data-driven and physical models. H5 is based on empirical models, data-driven, and physical models. Hu [2] et al. proposed the integration of a data-driven prediction algorithm and concluded that the prediction accuracy of the integrated algorithm is higher than that of a single algorithm. Li [3] et al. proposed a prediction method based on integrated learning. They considered the impact of degradation on RUL and made life predictions for the main bearings and prime movers of the ship. The results show that it is better than traditional integrated learning prediction methods.

Nowadays, there have been certain research achievements in the life prediction of ship propulsion devices in China. However, for the research and application of various hybrid methods to improve the accuracy of prediction, the control of various uncertain events needs further development. This article
mainly adopts a data-driven method to carry out research on the prediction of the remaining life of the propulsion device. We use the CMAPSS data set and apply the virtual health index method to preprocess the data. The engine performance degradation curve is constructed based on the theory of support regression machine, and we then combine the method based on similarity interpolation to realize the life prediction.

2. Supporting Vector Machine
Support vector machine (SVM) is a kind of binary classification model. Its basic model is a linear classifier with the largest interval in the feature space. It is a good way to realize the strategy of minimizing structural risks. When using SVM to solve the classification problem, we find a hyperplane that can accurately classify and maximize the classification interval. In addition, we require as few points as possible within the interval. In SVR, the concept of interval still exists, but we want to define an interval so that there are as many points as possible in this interval, so as to achieve data fitting, and take the middle hyperplane as the fitting result.

3. Performance parameter pretreatment of ship propulsion

3.1 Data sources
The data in this paper is obtained through the CMAPSS engine simulation model. The data includes the engine unit, the number of engine operating cycles, 3 operating parameters, and 21 sensor parameters. The entire data set is divided into an exercise set and a test set. The data in the exercise set reflects the life cycle of the engine unit from start to failure, while the data in the test set cuts off a cycle before the engine fails. This includes the parameters of the engine unit from the initial state to the cycle.

The three operating parameters in the data set represent the marine environment, match number, and sea temperature, respectively. The three operating parameters have a great influence on the performance of the propulsion device, and the changing trends of the parameters of the various sensors in different operating environments are also different. These 3 operating parameters can be combined into 6 operating environments. According to our operating environment, our data set is divided into 4 groups of practice sets and test sets. Each practice set and test set contains a part of the propulsion unit, as shown in Table 1.

| Data set name | FD001 | FD002 | FD003 | FD004 |
|---------------|-------|-------|-------|-------|
| Number of Exercise set units | 100   | 260   | 100   | 249   |
| Number of test set units | 100   | 259   | 100   | 248   |
| Operating environment | 1     | 6     | 1     | 6     |

Table 1 Data set classification

The physical meaning of 21 sensor parameters in the data set is shown in Table 2.

| symbol | significance | unit |
|--------|--------------|------|
| T2     | Total inlet temperature | R    |
| T24    | Total temperature at LPC outlet | R    |
| T30    | Total temperature of HPC outlet | R    |
### 3.2 Data collection

This forecast uses the FD001 data set for forecasting. The practice set and test set each contain the parameter values of 100 propulsion units. **Figure 1** shows part of the data in the exercise set. The first column represents the propulsion unit, and the second column represents the propulsion unit cycle number. Columns 3, 4 and 5 indicate 3 operating parameters. The next 21 columns represent 21 sensor parameters. The test set data is similar to the exercise set, but only data to a certain cycle.

![Figure 1](image)

The number of cycles in the data represents the life of the propulsion unit. The practice set includes the life of the propulsion unit and is used to practice the prediction model. The test set contains part of the life sequence, which can be used to predict its remaining life. In the 21-sensor data, we can see that
there is no obvious change trend of the parameters of many sensors with the cycle, which is not sensitive to the degradation of the performance of the propulsion device. In order to make the prediction model more accurate, we must choose a sensor that can reflect the performance degradation trend well. Based on literature [4], seven sensor parameters (2, 3, 4, 7, 11, 12, 15 in the 21-column sensor data) were selected for this design.

3.3 Data preprocessing
As mentioned above, 7-dimensional data is extracted from 21-dimensional data, which greatly reduces the dimension. However, it is still very complicated to deal with it. In addition, in order to be able to predict the life of the propulsion device well, we need to extract the health status of the propulsion device from a large amount of sensor data. We adopted the Virtual Health Index (VHI) method mentioned in Hu's research to characterize the health status of the propulsion unit. This method can convert multi-dimensional data into a one-dimensional VHI through a linear conversion method. The specific introduction of the virtual health index method is as follows.

We construct two data sets $Q_0$ and $Q_1$, which represent the failure data and health data of the propulsion system. And the two data sets are the matrix of $M_0 \times D$ and $M_1 \times D$ ($M_0$. $M_1$ represents the size of the data set; D represents the dimension of the data set). A matrix T can be constructed from these two matrices.

$$T \equiv (Q^T Q)^{-1} Q^T S_{off}$$

The matrix T is a VHI that can convert multi-dimensional sensor data into one-dimensional. $Q = [Q_0 ; Q_1]$ $S_{off} = [S_0 , S_1]^T$ is the formula ($S_0$ is a zero vector of $1 \times M_0$; $S_1$ is a unit vector of $1 \times M_1$). For the data of the practice set or the test set, it can be converted into a one-dimensional VHI through the matrix T, which is $H = Q_{off} T$ or $H = Q_{on} T$ (H is VHI; $Q_{off}$ represents the practice set for offline practice. $Q_{on}$ represents the test set for online prediction). H will be a list of vectors of size $Q_{off}$ or $Q_{on}$. VHI can also be expressed as $h(t_i), i = 1, ..., M_{off}$ or $i = 1, ..., M_{on}$, and the value varies from about 0 to 1.

Through the VHI method, we can establish the health background knowledge of the propulsion unit. In addition, we can convert multi-dimensional data into one-dimensional data to pave the way for subsequent offline practice prediction models and online predictions. In the 7-dimensional data selected in this design, the $Q_0$ failure data selects the sensor data of $0 \leq L \leq 4$ (L represents the remaining life of each propulsion unit, which is the sensor data when the remaining number of cycles of each unit is between 0 and 4) . $Q_1$ health data selected $L > 300$ sensor data.

4. Offline exercises for prediction models

4.1 Data collection
We extract each propulsion unit in the exercise set and test set through the Matlab program. The two data sets get an array of 100 propulsion units, respectively, and its size of the array related to the number of propulsion unit cycles and D dimension. Figure 2 is an array of partial cells extracted from the exercise set.
This paper uses the method of K-fold cross-validation to verify the performance of the model. In the offline practice phase, we divided the practice set into ten subsets, and each subgroup contains similar or identical propulsion unit units (in this case, 10). Nine of these subsets are practiced to obtain the model, then the remaining subset is used for prediction verification, and each subgroup must be verified once. The exercise set can be directly used to predict the remaining life, and the data in the exercise set still includes the full life. Therefore, we still need to randomly truncate each propulsion unit’s cyclic sequence in the practice set so that it can be divided into two parts. The former part is used for prediction, while the latter part is the real result. This is the data collection process through Matlab.

4.2 Establish performance degradation curve based on SVM

We mentioned that the VHI method is used to preprocess the data and the unit construction of offline practice \((h(t_i), i = 1, \ldots, M_{off})\), see chapter 3.3. The various parameters of each engine unit are attributed to a data set \(\{t_i, h_i\}\) \((t_i\) representing the time variable, the number of cycles). The various parameters of each engine unit are attributed to a data set (representing the time variable, the number of cycles). We construct predictive health degradation curves for each offline group.

\[
h(t) = \sum_{i=1}^{M} (\alpha_i - \alpha_i^*) k(t, t_i) + b
\]  

The regression equation selects the Gaussian kernel function for the kernel function. According to hu, regularization function \(C = 10\), insensitive loss function \(\varepsilon = 0.01\), and kernel function parameters \(g = 30\) can obtain good regression results. Figure 3 is the fitting curve established by the SVM regression method.
The blue dots represent each propulsion unit’s data in the exercise set, and the abscissa represents the adjusted number of cycles (for example, -250 indicates that there are still 250 cycles fail). The red line is the VHI curve obtained by the support vector regression.

4.3 Life prediction based on similarity interpolation

Reference [5] mentioned the prediction method based on similarity. The general principle is data from multiple propulsion units in the same system, thereby creating a reference library of performance degradation patterns. The general principle is data from multiple propulsion units in the same system, thereby creating a reference library of performance degradation patterns. When predicting the test unit, the test unit’s data can match the patterns in the reference library, and the actual service life of the matched group can be used as the basis of the estimation. This design adopts the similarity-based interpolation method mentioned in [5]. It uses the VHI constructed by the practice data as a reference library. In doing so, we can construct VHI on the prediction data to match. The life prediction process is divided into two steps.

1. Identify the initial health status of the predicted test unit.
2. Use similarity interpolation to predict the remaining life.

4.3.1 Identifying initial health status of the test unit

Different test units have different initial health indexes. If we want to make accurate life predictions, it is necessary to determine the test unit’s initial health accurately.

The VHI curve \( (h_p^t) \) of the exercise set unit was obtained by the SVR method above. According to the curve, an optimization function can be obtained and used to determine the number of health states on a time scale. The optimization formula is as follows.

\[
\text{Min } SSD = \sum_{j=1}^{M} \left( h_j^t(t_j) - h_p^t(t_j + T_0) \right)^2
\]

subject to \( T_0 \in [0, TS - \Delta t] \)

\( h_j^t(t_j), \ h_p^t(t_j) \) represent the test unit at the moment (in the offline practice stage, we use a subset of the practice set) and the VHI of the practice unit. \( \Delta t \) represents the number of cycles of the test unit, and TS represents the number of cycles of the practice unit (the full cycle life). When \( T_0 \) is determined, we can calculate the estimated life of the test unit according to the practice unit’s VHI curve.

\[
L^p = TS - \Delta t - T_0
\]

We repeat the optimal fitting process for \( K \) VHI curves of \( K \) different practice units, thereby giving \( K \) RUL estimates \( (L^p_i, i = 1, ..., K) \).

4.3.2 Life prediction based on similarity interpolation

The final predicted life is a linear interpolation function. The predicted remaining service life of a test unit shows that:

\[
L = \frac{1}{W} \sum_{i=1}^{k} \left( W_i \cdot L_i^p \right), W = \sum_{i=1}^{k} W_i
\]

\( W_i \) represents the \( i \)th similarity weight, and \( W_i = (SSD_i)^{-1} \).
4.3.3 Practice set prediction results

We adopt a 10x CV approach, and the prediction results of the exercise set are shown in Figure 4. Target in the SVM diagram means that the real remaining life is obtained by randomly cutting off the 100 exercises set propulsion unit before. The output represents the predicted remaining service life of 100 exercise set units used as test units. It can be seen from the figure that the two curves fit very well.

RMSE is the root mean square error and is the main measure of prediction accuracy. Calculation formula:

\[
RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (RUL_{true,i} - RUL_{predicted,i})^2}
\]

The predicted RMSE value this time is shown in Figure 5. The RMSE = 24.2778, and the accuracy is very good. Moreover, error represents the difference between the predicted value and the true value. At present, the parameters of this prediction model are more appropriate, so the prediction results are already more accurate. Eventually, it can be used for online prediction.
5. Online prediction and result analysis

5.1 Online prediction

To verify the robustness of the model and its accuracy after its promotion, we need to use the test set mentioned in 3.2 to predict the lifetime of the propulsion unit. The entire forecasting process is roughly the same, and the test set instead generates the CV test unit in the offline practice. We also need to predict each propulsion unit's remaining life in the test set one by one. Then we compare with the real remaining life of each propulsion unit in the test set. As a result, the data set RUL_FD001 provides the real remaining life of the test set. Figure 6 is the resulting graph predicted by the test set, and its two curves fit well. Figure 7 is the error curve of this prediction, RMSE=25.017.

![Figure 6 Test set prediction results.](image6)

![Figure 7 Test Set MSE](image7)

5.2 Result analysis

The calculation formula of the asymmetric index Score of the prediction result and the relative error MP is as follows:
\[ d_i = RUL_{\text{true},i} - RUL_{\text{predicted},i}, i = 1, ..., N \]

\[ MP = \frac{1}{N} \sum_{i=1}^{N} \frac{d_i}{RUL_{\text{true},i}} \]

\[ Score = \sum_{i=1}^{N} \left( e^{-\frac{d_i}{13}} - 1 \right), di < 0 \]

\[ Score = \sum_{i=1}^{N} \left( e^{\frac{d_i}{10}} - 1 \right), di > 0 \]

The RMSE together with the prediction results of the two data sets shown in Table 3

| RMSE  | SCORE   | MP      |
|-------|---------|---------|
| Exercise set | 24.2778 | 484.0628 | -0.0141 |
| Test set    | 25.0127 | 6.3619  | -0.0194 |

We found that the accuracy of the prediction model is very stable, and has good robustness and generalization through comparison several prediction performance indicators of the training set with the test set. We also compare and analyze the error graph and the predicted result graph. It can be found that the two curves in the prediction result fit well when the number of remaining lives is small. However, the closer the propulsion unit in error is to 0, the lower the remaining life. Therefore, when the prediction unit is close to failure, the prediction accuracy will be higher.

6. Conclusion
This article mainly builds the life prediction model of the ship propulsion device. We not only used Matlab software but also adopted the CMAPSS data set. Finally, the data is analyzed through the virtual health index method. In addition, this paper establishes the propulsion device's performance degradation curve based on the support vector regression method. Then, we create the prediction model by the method based on similarity interpolation. We found that the prediction model based on the support vector has good accuracy and robustness through the prediction verification of the data set. Further, as the prediction unit's remaining service life decreases, the prediction accuracy also shows an upward trend. In general, the purpose of this article is to predict the remaining service life of the propulsion unit to carry out corresponding health management, that is, preventive maintenance and repair.

References
[1] Liao L, Köttig F.(2014) Review of hybrid prognostics approaches for remaining useful life prediction of engineered systems, and an application to battery life prediction. IEEE Trans Reliab;63(1):191–207.
[2] Hu C, Youn B, Wang P, Yoon J. (2012) Ensemble of data-driven prognostic algorithms for robust prediction of remaining useful life. Reliab Eng Syst Saf;103:120–35.
[3] Zhixiong Li, Dazhong Wu, Chao Hu, Janis Terpenny. (2019) An ensemble learning-based prognostic approach with degradation-dependent weights for remaining useful life prediction. Reliab Eng Syst Saf;184:110–122.
[4] Wang T, Yu J, Siegel D, and Lee J., 2008, A similarity-based prognostics approach for remaining useful life estimation of engineered systems, International Conference on Prognostics and Health Management, Denver, CO, Oct 6–9.

[5] Zhixiong Li, Dazhong Wu, Chao Hu, Janis Terpenny. (2019) An ensemble learning-based prognostic approach with degradation-dependent weights for remaining useful life prediction. Reliab Eng Syst Saf;184:110–122.