Review of Common Evaluation Methods for Life Prediction Models

Yuan Fangcheng*

School of Reliability and System Engineering, Beihang University, Beijing, 100191, China

Abstract: As one of the core parts of Prognostics and Health Management (PHM) technologies, residual useful life (RUL) prediction is a very important concept in decision making and contingency mitigation. With life prediction models, researchers could obtain the prediction RUL of different objects. However, since sometimes there will be several available prediction models to be chosen, evaluation methods or selection methods) for life prediction models should be proposed to help choosing models that suit for certain objects. The most important factors that affect the performance of prediction models include prediction accuracy, data fitness, model complexity and parameter sensitivity. This paper presents some common evaluation methods for life prediction models that have already been used in this area.

Keywords: Evaluation, model complexity, parameter sensitivity, prediction accuracy, RUL prediction.

1. INTRODUCTION

Today's manufacturers face strong pressure on maintaining and supporting their complex, intelligent products since it becomes harder and harder to make right decisions. The high reliability and long life-circle products nowadays are really hard to fail, however, it will be a great disaster if a complex engineering system shut down suddenly. Predicting the precise failure time becomes more and more important. Therefore, Prognostics and Health Assessment (PHM) technology is proposed to solve these problems from the beginning to the end.

Prognostics and Health Assessment (PHM) technology mainly consists data acquisition, fault detection, fault diagnostics and prognostics, RUL prediction methods. It has already been widely induced in the area of aerospace and aviation industry, electronic engineering industry and even military industry. Prognostics is an engineering discipline focused on predicting the time at which a system or a component will no longer perform its intended function with certainty [1]. Prognostics predicts the future performance of a component by assessing the extent of deviation or degradation of a system from its expected normal operating conditions [2]. For the RUL prediction methods, prediction models is the most important part. Different models suit for different objects. There are mainly two kinds of models that have already been commonly used in this area: physical models and data-driven models.

Model selection methods have been developed on the condition if there are several prediction models that all work for the certain object. An evaluation guideline is needed to determine which prediction model is the best one based on the requirement. Some common factors that would have influence on the performance of prediction result include prediction accuracy, data fitness, model complexity, parameter sensitivity and so on.

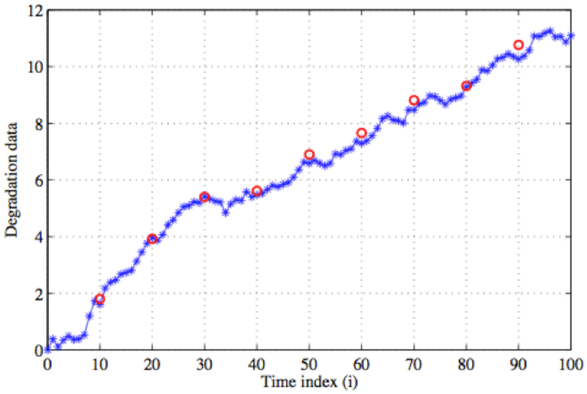

Fig. (1). Degradation data.

2. EVALUATION METHODS BASED ON PREDICTION ACCURACY

In some real applications, prediction results based on NN, ARIMA and SVM etc. are a series of certain values, so it's called point prediction. Fig. (**1**) presents a sample of the point prediction. Every red circle in the plot represents a certain prediction value while the blue line represents the real life distribution.

For this kind of prediction methods which have certain values as the final results, the prediction accuracy-based method is the best way to evaluate the effectiveness of the prediction model. The differences between real life distribution and prediction values could reflect the accuracy. Some performance indicators are shown below:

Root Mean Square Error (RMSE)

RMSE is a very common indicator for the model prediction accuracy evaluation. The smaller the RMSE is, the better performance the model has. Meanwhile, this value also

*Address correspondence to this author at the XueYuan Road No.37, HaiDian District, Beijing City, 100191, P.R. China; Tel: +86 18811434084; E-mail: yuanfc0530@126.com

1874-110X/14 2014 Bentham Open

could reflect the dispersion degree of the data. The smaller the RMSE is, the smaller the dispersion degree is.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (2.1)$$

Square Sum Error (SSE)

SEE is the summary of the square of the prediction error. It's more sensitive to the vibration of the relative error since the square enhances the error.

$$SE = \sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2 \text{ SSE} = \sum_{i=1}^{N}\left(y_i\text{-}\hat{y}_i\right)^2 \quad (2.2)$$

Mean Absolute Error (MAE)

The purpose of having absolute value here is to avoid the positive and negative error offset. This indicator has a great value on evaluating the model prediction accuracy. However, it can't reflect the minor changes of the prediction error. The sensitivity of the MAE could be improved by increasing prediction error.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right| \quad (2.3)$$

Percent Error (PE)

PE is a percentage value which is the ratio of the true value and the prediction value.

$$PE = \frac{y_i - \hat{y}_i}{y_i} \times 100\% \quad (2.4)$$

Mean Percent Error (MPE)

MPE is the average value of PE. It could be seen that the positive and negative offset makes the final error smaller than it should be.

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{y_i - \hat{y}_i}{y_i} \times 100\% \quad (2.5)$$

Mean Absolute Percent Error (MAPE) The absolute value avoids the offset problem so the MAPE is a reasonable indicator.

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{y_i} \times 100\% \quad (2.6)$$

where, y_i donates the true degradation value that has been monitored; $\hat{y}_i$ donates the prediction value; N is the number of all observation points.

The former three indicators above use standard statistical metrics, the smaller the indicator is, the better performance the model has. However, since there is not a unified guideline to judge how small the indicator is, an effective result could not be obtained when we evaluate different models. The later three indicators use relative guideline. These indicators all have a unified guideline, which is the ratio of the true value and the prediction value so that they are not affected by the dimensions.

Most of the papers about life prediction models used some of these indicators as their evaluation guidelines. For the same object, if the indicators of the new model are smaller than the old ones, it means the model they proposed has better performance on that situation. It is noted that a prediction method is not valuable if it only has high accuracy on some certain points. Actually, an efficient prediction model should have high prediction accuracy on all points.

In some other situations, point prediction can not describe the uncertainty of the prediction, which is the natural character of prediction method. In a sense, a prediction model without considering the uncertainty is meaningless. The prediction result of every single point should be an interval. Commonly, the uncertainty could be derived from model structure or Monte-Carlo Bootstrap method. The research about the prediction uncertainty have been discussed in some papers [3-5]. Fig. (**2**) presents a sample of prediction model of uncertainty prediction. The red circle is the mean value of prediction while it also shows the interval of every prediction point.

Saxena A, *et al.* [6] presented several indicators for evaluation of life prediction models. The $\alpha - \beta$ indicator is the most widely used one. It's could be defined as:

$$(1 - \alpha)(t_{EOL} - t) \leq F(t) \leq (1 + \alpha)(t_{EOL} - t) \quad (2.7)$$

$$t = t_p + \lambda\left(t_{EOL} - t_p\right) \quad (2.8)$$

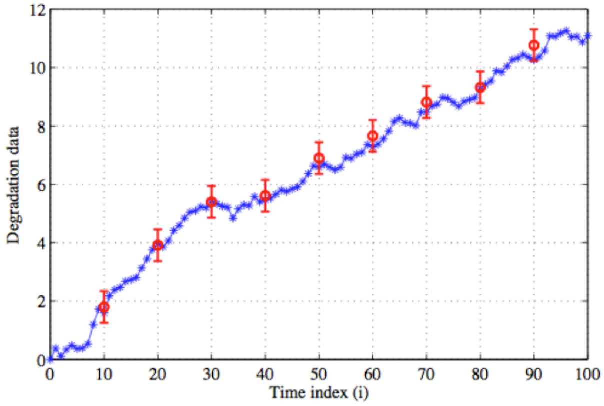

Fig. (2). Example of uncertainty prediction.

where, t_{EOL} donates the real life of sample l; t_p donates the first observation point; λ is the window controller and α is precision controller. Fig. (**3**) shows this indicator when α = 0.2.

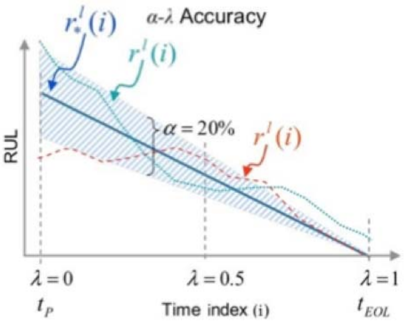

Fig. (3). $\alpha - \beta$ indicator when α = 0.2

In Fig. (**3**), $r_*^l(i)$ means the real RUL of sample l at time t_i while $r^l(i)$ donates the prediction value. We can see that this indicator allows larger error when the sample in its early time and the later the smaller it allows. In real application, prediction value larger than real RUL always leads to worse results. Base on this, T.Y. Wang [7] presents another rule in their life prediction method. It's defined as below:

$$S(i) = \frac{1}{L}\sum_{l=1}^{L} S^l(i) \tag{2.9}$$

$$S^l(i) = \begin{cases} \exp\left(-\frac{d^l(i)}{a_1}\right) - 1, d^l(i) < 0 \\ \exp\left(\frac{d^l(i)}{a_2}\right) - 1, d^l(i) \geq 0 \end{cases} \tag{2.10}$$

where, $a_1 > a_2 > 0$, $d^l(i) = r_*^l(i) - r^l(i)$. It is obvious that this indicator gives more punishment when the prediction is larger that real value.

3. INFORMATION CRITERION

Information Criterion is a model selection approach. It's a measure of the relative quality of a statistical model for a given set of data. Information Criterion (IC) deals with the trade-off between the goodness of fit of the model and the complexity of the model. Both of these two indicators could be used to evaluate the prediction model.

Akaike *et al.* [8] proposed the first information criterion method called Akaike Information Criterion in 1974. On the basis of AIC, N. Sugiura [9] proposed another method named as AICc in 1978. For any statistical model, the AIC value is defined as:

$$AIC = 2k - 2\ln(L) \tag{3.1}$$

The AICc value is defined as:

$$AICc = AIC + \frac{2k(k+1)}{n-k-1} \tag{3.2}$$

where, k donates the number of the parameters in the model; n donates the sample size and L is the maximized value of the likelihood function for the model. Thus, the AICc is AIC with a greater penalty for extra parameters in the statistical model.

Also in 1978, the Bayesian Information Criterion (BIC) was developed by G.E. Schwarz [10], who gave a Bayesian argument for adopting it. The BIC could be describe as:

$$BIC = -2\ln L + k(\ln(n) + \ln(2\pi)) \tag{3.3}$$

when n is large, the BIC could be approximately equal to:

$$BIC \approx -2\ln L + k\ln(n) \tag{3.4}$$

The Deviance Information Criterion (DIC) is a hierarchical modeling generalization of the AIC and BIC.

Define the deviance and expectation as:

$$D(\theta) = -2\log\big(p(y|\theta)\big) + C \tag{3.5}$$

$$\overline{D} = E^\theta(D(\theta)) \tag{3.6}$$

where y are the data, θ are the unknown parameters of the model and $p(y|\theta)$ is the likelihood function. C is a constant value that cancels out in all calculations that compare different models, which does not need to be known. There are two calculations in common usage for the effective number of parameters of the model, described by Spiegelhalter *et al.* [11] and Gelman *el al.* [12] respectively:

$$p_D = \overline{D} - D(\bar{\theta}) \tag{3.7}$$

$$p_D = \frac{1}{2}\widehat{VAR}(D(\theta)) \tag{3.8}$$

Then, the DIC is described as:

$$DIC = p_D + \overline{D} = D(\bar{\theta}) + 2p_D \tag{3.9}$$

The Focused Information Criterion (FIC), unlike most model selection strategies, does not attempt to assess the overall fit of candidate models but focuses attention directly on the parameter of primary interest with statistical analysis for which competing models lead to different estimates for a certain model. It was first developed by Gerda Claeskens *et al.* [13] and Nils Lid Hjort *et al.* [14] in two discussion articles.

Take Akaike Information Criterion as an example to illustrate how to apply the information criterion into real application for model selection. Starting with a set of candidate prediction models, and then find the corresponding AIC values of each model. Since there will almost always be information loss due to using on of the candidate models to represent the real model, which generates the data, the one model that could minimize the loss will the best one among all others. Because of different requirement, the model cannot be chosen with certainty but it should minimize the estimated information loss in this single situation.

Donate the AIC values of all candidate models as AIC$_1$, AIC$_2$, AIC$_3$,…,AIC$_{N.}$ AIC$_{min}$ is the minimum of all values. Then the relative probability that the *i*th candidate model minimizes the estimated information loss $f(i)$ can be described as below:

$$f(i) = \exp\left((AIC_{min} - AIC_i)/2\right) \tag{3.10}$$

Although the model with the minimum AIC value is the best one among all others, it still should be considered if there are any other $f(i)$ values are very close to 1. The closer the $f(i)$ value to 1, the fewer information loss that the model could have. Then there are three choices:

Gather more data to distinguish the top models more clearly;

Simply conclude that the data is insufficient to support selecting one model among the top models.

Take a weighted average of the top models based on the $f(i)$ value of each model. Then do statistical inference based on the weighted multiple models [15].

It should be noted that if all candidate models have the same numbers of unknown parameters, the result of using AIC method might at first be very similar to using likelihood-ratio test.

4. PARAMETER SENSITIVITY ANALYSIS

Parameter sensitivity means that the model results can be highly correlated with an input parameter so that small changes in the parameter could result in significant changes in the output [16]. Crick *et al.* [17] made a distinction between important parameter, whose uncertainty contributes substantially to the uncertainty in assessment results, and sensitive parameter, which have a significant influence on assessment results. For life prediction models, it will be much harder to get a good result if the model has too many high sensitive parameters.

One method for parameter sensitivity is called partial sensitivity analysis. It tests the changing of the model output by changing a single parameter value. Since it only analyzes one parameter, this method is relatively easy to accomplish. The partial parameter sensitivity of x_i in model $F_j(X)$ at X_k could be obtained from:

$$S_{ji} = \frac{\partial F_j(X)}{\partial x_i}\bigg| X = X_k \qquad (4.1)$$

However, partial sensitivity analysis does not consider the output changing caused by the parameters interactions since most models have more than one parameter. The natural limitation of this method makes it not suit for many prediction models that have many parameters. The global sensitivity analysis method could solve this problem. The main differences between partial method and global method are:

Global method considers the effect of different value of one parameter for sensitivity analysis, which is more reasonable.

Global method can obtain the integrated sensitivity of all parameters.

Since the global sensitivity analysis is more practical in real applications, many methods have been proposed. Mckay *et al.* [18] proposed a multiple regression method; M.D. Morris [19] named his method as Morris method; Cukier *et al.* [20] proposed a Fourier Amplitude Sensitivity Test (FAST) method.

CONCLUSION

RUL prediction has already had more and more attention from the academia and engineering. A good evaluation method for these prediction models could help people find their most suitable models faster. This paper presents some common evaluation methods that had already been used in some papers. However, they all just consider only one factor that could have influence on prediction model performance. Therefore, a method that could consider all factors is still needed to be built in the future.

CONFLICT OF INTEREST

The author confirms that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] Vachtsevanos, Lewis, Roemer, Hess, and Wu, “Intelligent fault Diagnosis and Prognosis for Engineering Systems”, *Wiley*, USA, 2006.

[2] Pecht, and Michael, “Prognostics and Health Management of Electronics, *Wiley*, USA, 2008.

[3] L. Tang, and G. L. Kacprzynski, “Methodologies for uncertainty management in prognostics”, Proceedings of the IEEE Aerospace Conference, 2009.

[4] S. Sankararaman, and M. Daigle, “Analytical algorithms to quantify the uncertainty in remaining useful life prediction”, Proceedings of the IEEE Aerospace Conference, 2013.

[5] P. Baraldi, and F. Mangili, “Investigation of uncertainty treatment capability of model-based and data-driven prognostic methods using simulated data”, *Reliability Engineering and System Safety*, vol. 112, pp. 94-108, 2013.

[6] A. Saxena, and J. Celaya, “Metrics for offline evaluation of prognostics performance”, *International Journal of Prognostics and Health Management*, vol. 1, pp. 1-20, 2010.

[7] T. Y. Wang, “A similarity-based prognostics approach for remaining useful life estimation of engineered systems”, Proceedings of the IEEE Conference on Prognostics and Health Management, 2008.

[8] Akaike, and Hirotugu, “A new look at the statistical model identification”, *IEEE Transactions on Automatic Control*, vol. 19, pp. 716-723, Dec 1974.

[9] N. Sugiura, “Further analysis of the data by Akaike’s information criterion of model fitting”, *Communications in Statistics – Theory and Methods*, vol. A7, pp. 13-26, 1978.

[10] G. E. Schwarz, “Estimating the dimension of a model”, *Annals of Statistics*, vol. 6, pp. 461-464, Mar 1978.

[11] Spiegelhalter, and J. David, “Bayesian measures of model complexity and fit”, *Journal of the Royal Statistical Society*, vol. B64, pp. 583-639, Oct 2002.

[12] A. Gelman, J. B. Carlin, H. S. Sterm and D. B. Rubin, “Bayesian Data Analysis: Second Edition”, *CRC Press*, USA, 2004.

[13] G. Claeskens, and N. L. Hjort, “The focused information criterion”, *Journal of the American Statistical Association*, vol. 98, pp. 879-899, 2003.

[14] N. L. Hjort, and G. Claeskens, “Frequentist model average estimators”, *Journal of the American Statistical Association*, vol. 98, pp. 900-916, 2003.

[15] K. P. Burnham, and D. R. Anderson, “Model Selection and Multimodel inference: A practical Information – Theoretic Approach: second edition”, *Springer Science & Business Media*, USA, 2002.

[16] D. M. Hamby, “A Review of Techniques for Parameter Sensitivity Analysis of Environment Models” *Environment Monitoring and Assessment*, vol. 32, pp. 135-154, Sep 1994.

[17] M. J. Crick, and M. D. Hill, “The Roll of Sensitivity Analysis in Assessing Uncertainty”, Proceedings of an NEA Workshop on Uncertainty Analysis for Performance Assessments of Radioactive Waste Disposal Systems, Paris, PP. 1-258, 1987.

[18] M. D. Mackay, and R. J. Beckman, “A comparison of three methods for selecting values of input variables in the analysis of output from a computer code”, *Technometrics*, vol. 21, pp. 239-245, 1979.

[19] M. D. Morris, “Factorial sampling plans for preliminary computational experiments”, *Technometrics*, vol. 33, pp. 161-174, 1991.

[20] R. I. Cukier, C. M. Fortuin and K. E. Shuler, “Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients”, *Journal of Chemical Physics*, vol. 59, pp. 3873-3878, Oct 1973.

Received: September 16, 2014 Revised: December 23, 2014 Accepted: December 31, 2014