基于随机森林算法的航空发动机振动趋势预测

王孝军，刘德虎
(中国航发四川燃气涡轮研究院，成都 610500)

摘　要：提出了基于随机森林算法的航空发动机振动趋势预测模型。阐述了随机森林算法的基本理论和计算步骤，采用C-C法计算了延迟时间和嵌入维数，对一维时间序列进行了相空间重构，并在此基础上建立了随机森林算法的预测模型。应用发动机振动试验数据进行了振动预测，并与利用相同训练数据建立的支持向量机预测模型的预测结果进行对比。结果表明，与支持向量机模型相比，随机森林算法预测模型的预测精度更高，泛化能力更强，操作方便，且计算效率更高。

关键词：航空发动机；振动；随机森林算法；趋势预测；相空间重构；支持向量机
中图分类号：V263.5　　**文献标识码：**A　　**文章编号：**1672-2620 (2020) 02-0039-05

Aero-engine vibration trend prediction based on random forest algorithm

WANG Xiao-jun, LIU De-hu
(AECC Sichuan Gas Turbine Establishment, Chengdu 610500, China)

Abstract: A random forest model was proposed for aero-engine vibration prediction. Firstly, the basic principles and the process of random forest algorithm was introduced, and then C-C method was applied to calculate the delay time and embedding dimension, the time series was rebuilt by phase space reconstruction. Finally, the random forest prediction model was established, and used to forecast aero-engine vibration by applying test data. Based on the same data, the prediction results were compared with that of support vector machine. The results indicate that the random forest model has higher precision and better generalization. Compared with support vector machine prediction model, the random forest prediction model is much easier to operate, and has higher computational efficiency.

Key words: aero-engine; vibration; random forest algorithm; trend prediction; phase space reconstruction; support vector machine

1　引言

航空发动机转动部件故障通常与振动有关。因此，对发动机振动趋势进行预测，及早发现潜在问题，对发动机实现安全、可靠的工作尤为重要。对于振动预测，学者们提出了多种方法，主要有：时间序列分析[1-2]、神经网络[3]、支持向量机[4-5]等方法。时间序列分析方法适用于线性系统的短期预测，对于航空发动机振动趋势预测这类非线性问题不太适用；神经网络方法能拟合任意非线性函数并具有一定的泛化能力，但实际运用时训练样本的选择对神经网络模型的影响较大，且神经网络容易出现过拟合现象[6]；支持向量机方法虽然泛化能力较强，但存在获取最优参数计算量较大的问题[7]。

为克服单一算法的缺陷，集成算法得到了学者的深入研究。集成算法通过构建多个不同的预测模型，然后根据某种规则组合输出预测结果，目前理论上已被证明可以获取良好的学习效果[8-9]。随机森林(RF)算法是一种典型的集成算法，采用了 Random

收稿日期：2019-08-13
作者简介：王孝军(1986-)，男，四川苍溪人，工程师，硕士，主要从事航空发动机总体性能技术研究。

subspace method和Bootstrap aggregating两大随机思想[10],能较好地避免过拟合现象,且随机森林算法易于实现,训练速度快,能有效分析非线性数据,预测精度高。目前,随机森林算法已在电力负荷预测[11]、交通预测[12]及水文预测[13]等领域得到了广泛的研究和应用,但针对航空发动机状态预测方面的研究较少。为此,本文利用某发动机振动测量样本数据,建立随机森林算法预测模型,对航空发动机振动趋势进行预测,旨在为航空发动机振动趋势预测分析提供新的方法。

2 随机森林算法原理

2.1 算法概述

随机森林算法是一种集成算法,通常采用分类回归树(CART)作为基元学习器,通过建立许多相互之间没有关联的树,得到一个组合学习模型。最终输出的预测值对于分类问题采用多数投票法,对于回归问题则是整个森林预测值取平均值。由于随机森林算法在构建每棵决策树时引入袋装法和特征子空间法两大随机策略,使随机森林算法集成了各决策树的分类回归结果,抵消了部分随机误差,对异常值和噪声具有很好的容忍度[14-15],其计算步骤详见文献[15]。

2.2 泛化误差分析

随机森林泛化误差已经得到证明[16],本文仅作简单介绍。

给定 k 个决策树集合 $\{h(\boldsymbol{X},\boldsymbol{\theta}_k),k=1,2,\cdots n\}$,其中 $\boldsymbol{\theta}_k$ 是相互独立且同分布的随机向量,$h(\boldsymbol{X},\boldsymbol{\theta}_k)$ 是未减枝的CART树,最终结果由所有树的投票结果决定。定义随机函数的泛化误差:

$$PE^*=P_{X,Y}\left(K(\boldsymbol{X},Y)<0\right) \tag{1}$$

式中:$P_{X,Y}$ 为给定输入向量 $\boldsymbol{X}$ 的分类错误率函数,$K(\boldsymbol{X},Y)$ 为随机森林的边缘函数。当森林中树的数目较大时,可以用大数定律得到如下定理。

定理1 随着树的数目增加,对于所有随机变量 $\boldsymbol{\theta}$,PE^* 将收敛于:

$$P_{X,Y}\left(P_\theta\left(h(\boldsymbol{X},\boldsymbol{\theta})=Y\right)-\max_{j\neq Y}P_\theta\left(h(\boldsymbol{X},\boldsymbol{\theta})=j\right)<0\right) \tag{2}$$

式中:P_θ 为对于给定序列 $\boldsymbol{\theta}$ 分类错误。

该定理表明随着树的数目增加,随机森林的泛化误差趋于某一上界,而不会造成过拟合,这是随机森林的一个重要特点[17]。

定理2 定义随机森林的泛化误差上界:

$$PE^*\leqslant\frac{\bar{\rho}(1-s^2)}{s^2} \tag{3}$$

式中:$\bar{\rho}$ 为关系数的均值,s 为树的分类强度。

随着树的相关性的降低和单棵树强度的提高,随机森林的泛化误差上界将减小,其泛化误差将得到控制[7]。

2.3 袋外数据评估

在抽取样本子集时,原始数据集中每个样本未被抽取的概率为 $\left(1-\frac{1}{N}\right)^N$($N$ 为原始数据集样本总数),当 N 足够大时,$\left(1-\frac{1}{N}\right)^N$ 收敛于 $1/e\approx0.368$。因此总有约37%的数据样本未被选中,这些数据称为袋外数据(OOB)。随机森林算法采用在袋外数据上预测的残差均方进行回归效果评价,已证明OOB误差为无偏估计,是一种较好的泛化误差分析方法,可作为模型预测效果的验证,而不需要使用交叉验证的方式,提高了参数的调节效率[18]。

3 振动预测模型设计

3.1 时间序列相空间重构

试验数据来自文献[3],与其原始数据值不完全相等,旨在分析振动参数时间序列预测方法的适用性。振动原始数据为一维时间序列,通过基本的数据预处理后,数据样本共121个点,样本数较少。由于随机森林算法对于小样本问题的预测不理想,受文献[19]的启发,假定正常情况下发动机振动变化趋势为一渐进变化,提出通过建立等差数列内插值来构造大样本训练集。即在原始两个采样时刻之间等间隔增加虚拟采样点,从而扩增训练样本集。

为提高预测精度,需对一维振动时间序列进行相空间重构,充分挖掘数据间的关联关系,以获取尽可能多的信息。设原始序列 $\boldsymbol{X}=\{X_1,X_2,X_3,\cdots X_n\}$,通过重构后的具体形式为:

$$\boldsymbol{X}=\begin{bmatrix}X_1 & X_2 & X_3 & \cdots & X_m\\ X_2 & X_3 & X_4 & \cdots & X_{m+1}\\ X_3 & X_4 & X_5 & \cdots & X_{m+2}\\ \vdots & \vdots & \vdots & \cdots & \vdots\\ X_{n-m} & X_{n-m+1} & X_{n-m+2} & \cdots & X_{n-1}\end{bmatrix} \tag{4}$$

$$\boldsymbol{Y}=\left[X_{m+1},X_{m+2},X_{m+3},\cdots X_n\right]^{\mathrm{T}}$$

上式表示前 m 个振动参数信息能表达第 $m+1$ 个振动参数的信息。即对发动机当前时间点上振动状态的预测,主要依据的信息是发动机在之前 m 个

时间点上的振动趋势分布信息。

相空间重构时，合理选取嵌入维数 m 和时间延迟 τ 是关键。利用混沌理论中的C-C法可以同时估算出 τ 和延迟时间窗口 τ_w，两个参数间的关系为 $\tau_w=(m-1)\tau$ [20]。关联积分 $C(m,r,t)$，三个统计量 $\overline{S}(t)$、$\Delta\overline{S}(t)$ 和 $S_{cor}(t)$ 等参数详细的定义和计算方法见文献[20]。按照C-C方法的基本理论，$\Delta\overline{S}(t)$ 的第一个极小值对应于时间延迟，$S_{cor}(t)$ 的最小值对应于时间延迟窗口。经计算，时间延迟取5，时间延迟窗口取18，因此嵌入维数 m 可以取为5。

3.2　模型参数优化

随机森林算法的调节参数主要有决策树数目 n_{tree} 和决策树每次节点分割时随机选取的特征数量 m_{try}。n_{tree} 一般不少于100，要获得模型最佳性能需要进行调试获取最优 n_{tree}。查找最佳分割点时，选取的特征数量对模型有一定的影响。根据文献[21]，原始数据特征数目为 M，回归计算时建议取值为 $sqrt(M)$。以OOB误差最小为训练目标，通过计算得到最优的 n_{tree}。

其他参数设为默认值，仅 n_{tree} 变化时，OOB误差与决策树数目的关系如图1所示。可看出，OOB误差随着树的增多逐渐减少，并趋于稳定，在一个小区间范围内波动。经计算，n_{tree} 取340时OOB误差达到最小。n_{tree} 选定340，其他参数设为默认值，仅 m_{try} 变化时，在测试集样本的预测得分与特征参数量的关系如图2所示。可看出，m_{try} 取2时模型预测性能最好。

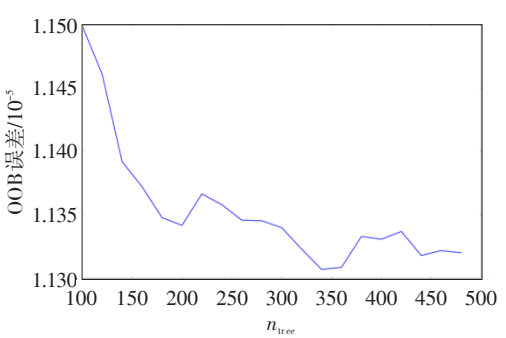

图1　决策树数目与袋外数据误差的关系

Fig.1　Relationship between n_{tree} and OOB error

3.3　模型性能评估指标

运用随机森林算法进行回归计算时对模型预测精度的评估，本文考虑均方误差（*MSE*）、平均绝对误差（*MAE*）以及预测性能得分（T_{score}）。*MSE* 和 *MAE* 越小，T_{score} 越大，说明模型精度越高，预测效果越好。各评价指标计算公式如下：

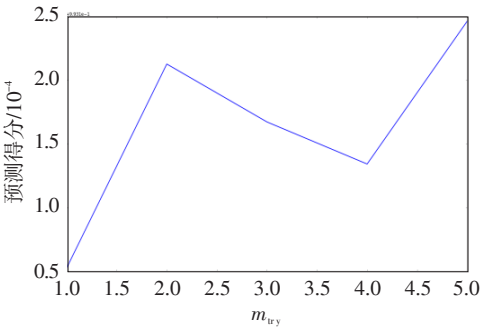

图2　测试集样本预测得分与特征数量的关系

Fig.2　Relationship between m_{try} and prediction score

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{5}$$

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i| \tag{6}$$

$$T_{score}=1-\frac{\sum_{T_{test}}(y_i-\hat{y}_i)^2}{(y_i-\bar{y}_i)^2} \tag{7}$$

式中：T_{test} 为测试集，y_i 为真实值，$\hat{y}_i$ 为预测值，$\bar{y}_i$ 为真实值的均值。

4　预测结果分析

运用上述讨论得到的随机森林算法预测模型，对发动机振动时间序列进行预测。将数据样本的90%用作训练样本，剩余的10%作为测试样本。预测结果见图3，测试集预测结果见表1，训练样本和测试样本的三个模型性能评估指标得分见表2。由预测结果可知，随机森林算法预测精度较高，泛化能力较强，很好地预测了发动机振动的变化趋势。

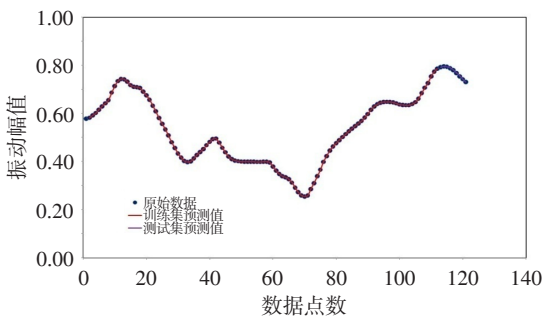

图3　随机森林算法模型预测结果与真实值的对比

Fig.3　Comparison between actual and predicted value by random forest algorithm model

为对比分析随机森林算法模型性能，用相同数据样本建立支持向量机(SVM)预测模型。支持向量

表1 测试集上预测值与真实值的对比

Table 1 Comparison between actual value and prediction value during test data

数据点	113	114	115	116
真实值	0.791 7	0.794 8	0.793 7	0.787 5
预测值	0.790 7	0.794 6	0.794 1	0.787 6
数据点	118	119	120	121
真实值	0.767 3	0.753 7	0.741 9	0.731 0
预测值	0.769 7	0.756 4	0.742 2	0.731 1

表2 随机森林算法模型评估指标分析

Table 2 Analysis of evaluation index of random forest algorithm model

	MSE	*MAE*	T_{score}
训练集	0.003 5	0.000 14	0.999
测试集	0.000 001 62	0.000 86	0.997

机模型核函数采用高斯函数，通过网格搜索寻优的方法获得模型最优的惩罚系数 C 、核参数 γ 、回归参数 ε ，支持向量机模型与随机森林算法模型预测结果的对比分别见表3和图4。从预测结果可知，两种模型均取得了满意的预测结果，在测试集上随机森林算法模型预测得分0.997，而支持向量机模型得分为0.987，说明随机森林算法的预测精度更高，泛化能力更强。另外，在训练时间上，支持向量机参数寻优步骤导致模型训练花费时间较长，整个参数优化过程时间比随机森林算法的多几倍。综合对比在模型设计及参数调试中的体验，随机森林算法模型综合能力更优异，操作更简单，更适合解决实际发动机振动预测问题。

表3 两种模型预测结果对比

Table 3 Comparison of the prediction results of two models

	MSE	*MAE*	T_{score}
RF	1.620×10^{-6}	0.000 86	0.997
SVM	6.666×10^{-6}	0.002 14	0.987

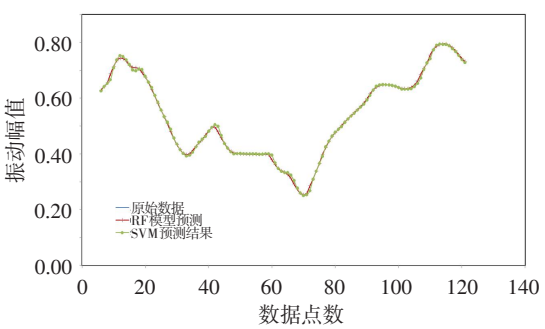

图4 两种模型预测结果

Fig.4 Prediction results of two models

5 结论

针对航空发动机振动趋势预测，构建了基于随机森林算法的振动预测模型，并应用发动机振动试数据验证了模型的适用性。结果表明，随机森林算法模型精度高、泛化能力强，能有效预测发动机振动趋势。此外，随机森林算法模型不需要进行数据归一化预处理，以及交叉验证等步骤，具有操作简单、调参容易、训练时间短等特点，尤其适用于大样本的发动机趋势预测分析。

参考文献：

[1] 黎瑜春. 组合模型在汽轮机转子振动状态预测中的应用[D]. 北京：华北电力大学，2012.

[2] Hai X, Wang X J, Jay L, et al. Engine health assessment and prediction using the group method of data handling and the method of match matrix: autoregressive moving average[C]//. ASME Turbo Expo 2007: Power of land, sea and air. 2007.

[3] 金向阳，林 琳，钟诗胜，等. 航空发动机振动趋势预测的过程神经网络法[J]. 振动、测试与诊断，2011，31(3)：331—334.

[4] Lv X S. Application of support vector machine with particle swarm optimization algorithm in blasting vibration prediction[C]//. ICEEE 2010 International Conference. 2010.

[5] 刘林刚，李学仁，陈永刚，等. 基于支持向量机的航空发动机振动预测模型研究[J]. 微计算机信息(测控自动化)，2008，24(6)：289—291.

[6] 孙智源. 基于过程神经网络集成的航空发动机性能衰退预测[D]. 哈尔滨：哈尔滨工业大学，2010.

[7] 吴潇雨，和敬涵，张 沛，等. 基于灰色投影改进随机森林算法的电力系统短期负荷预测[J]. 电力系统自动化，2015，39(12)：50—55.

[8] Schapire R E. The strength of weak learnability[J]. Machine Learning，1990，5(2)：197—227.

[9] 钟诗胜，雷 达. 一种可用于航空发动机健康状态预测的动态集成极端学习机模型[J]. 航空动力学报，2014，29(9)：2085—2090.

[10] 李威威，李春青，聂敬云，等. 膜生物反应器膜污染的随机森林预测模型[J]. 计算机应用，2015，35(S1)：135—137.

[11] 李婉华，陈 宏，郭 昆，等. 基于随机森林算法的用电负荷预测研究[J]. 计算机工程与应用，2016，52(23)：236—243.

[12] 程 政，陈贤富. 基于随机森林模型的短时交通预测方法[J]. 微型机与应用，2016，35(10)：46—49.

[13] 余胜男，陈元芳，顾圣华，等. 随机森林在降水量长期预报中的应用[J]. 南水北调与水利科技，2016，14(1)：78—

82.
[14] 董师师，黄哲学. 随机森林理论浅析[J]. 集成技术，2013，2(1)：1—7.
[15] 甄亿位，郝　敏，陆宝宏，等. 基于随机森林的中长期降水量预测模型研究[J]. 水电能源科学，2015，33(6)：6—10.
[16] Breiman L. Bagging prediction[J]. Machine Learning，1996，11(2)：123—140.
[17] 庄进发，罗　健，彭彦卿，等. 基于改进随机森林的故障诊断方法研究[J]. 计算机集成制造系统，2009，15(4)：777—785.
[18] 罗　浩，郭盛勇，包为民. 拱坝变形监测预报的随机森林模型及应用[J]. 南水北调与水利科技，2016，14(6)：116—121.
[19] 崔东文. 随机森林回归模型及其在污水排放量预测中的应用[J]. 供水技术，2014，8(1)：31—36.
[20] 商　强，杨兆升，李志林，等. 基于相空间重构和RELM的短时交通流量预测[J]. 华南理工大学学报(自然科学版)，2016，44(4)：109—114.
[21] Michael B. 预测分析核心算法：Python机器学习[M]. 沙赢，李　鹏，译. 北京：人民邮电出版社，2017.

(上接第24页)

温完成，需使用气化液氮和喷水强制冷却，且此过程需使用伺服系统来实现涡轮叶片在辐射热冲击区域和强制冷却区域的平移，伺服系统和强制冷却系统的介入和结束时间需根据热载荷条件确定。

参考文献：

[1] 关　鹏，艾延廷，王　志，等. 涡轮导向叶片热冲击数值模拟研究[J]. 推进技术，2016，37(10)：1938—1945.
[2] 徐庆泽，梁春华，孙广华，等. 国外航空涡扇发动机涡轮叶片热障涂层技术发展[J]. 航空发动机，2008，34(3)：52—56.
[3] Zaretsky E V，Litt J S，Hendricks R C，et al. Determination of turbine blade life from engine field data[J]. Journal of Propulsion and Power，2012，28(6)：1156—1167.
[4] Abu A O，Eshati S，Laskaridis P，et al. Aero-engine turbine blade life assessment using the Neu/Sehitoglu damage model[J]. International Journal of Fatigue，2014，61(4)：160—169.
[5] Lee J M，Song H W，Kim Y S，et al. Evaluation of thermal gradient mechanical fatigue characteristics of thermal barrier coating，considering the effects of thermally grown oxide[J]. International Journal of Precision Engineering and Manufacturing，2015，16(7)：1675—1679.
[6] 钱惠华，李　海，程　滔，等. 涡轮导向叶片热疲劳分析[J]. 航空动力学报，2003，18(2)：186—190.
[7] 骆剑霞，朱惠人，张宗卫. 某涡轮导向叶片换热实验与计算[J]. 航空动力学报，2014，29(3)：526—531.
[8] 段红燕，王小宏，张洹榕，等. 基于热力耦合计算的涡轮叶片疲劳/蠕变寿命预测[J]. 兰州理工大学学报，2017，43(4)：59—65.
[9] Ingram J，Gross L. Turbine blade thermal fatigue testing Pratt and Whitney aircraft hollow core blades[R]. NASA TM-86528，1985.
[10] Kerezsi B B，Kotousov A G，Price J W H. Experimental apparatus for thermal shock fatigue investigations[J]. International Journal of Pressure Vessels and Piping，2000，77(7)：425—434.
[11] Panda P K，Kannan T S，Dubois J，et al. Thermal shock and thermal fatigue study of alumina[J]. Journal of the European Ceramic Society，2002，22(13)：2187—2196.
[12] Liu Y C，He Y R，Yuan Z G，et al. Numerical and experimental study on thermal shock damage of CVD ZnS infrared window material[J]. Journal of Alloys and Compounds，2014，589：101—108.
[13] Vincent L，Poncelet M，Roux S，et al. Experimental facility for high cycle thermal fatigue tests using laser shocks[J]. Procedia Engineering，2013，66：669—675.
[14] Song H W，Yu G，Tan J S，et al. Thermal fatigue on pistons induced by shaped high power laser. Part I：Experimental study of transient temperature field and temperature oscillation[J]. International Journal of Heat and Mass Transfer，2008，31(4)：757—767.
[15] 李成刚，柳恩杰，郝　兵，等. 热机械疲劳试验器的研制[J]. 航空发动机，2004，30(3)：8—10.
[16] 张东明，柳恩杰. 航空发动机涡轮叶片高温振动疲劳试验的新方法[J]. 航空发动机，2005，31(1)：18—21.
[17] 王洪斌. 涡轮叶片热/机械复合疲劳试验方法研究[J]. 航空发动机，2007，33(2)：7—11.
[18] Wang R Q，Jing F L，Hu D Y. In-phase thermal-mechanical fatigue investigation on hollow single crystal turbine blades[J]. Chinese Journal of Aeronautics，2013，26(6)：1409—1414.
[19] Wang R Q，Jiang K H，Jing F L，et al. Thermomechanical fatigue failure investigation on a single crystal nickel superalloy turbine blade[J]. Engineering Failure Analysis，2016，66，284—295.
[20] 张　珏，张伯良. 结构热试验技术[M]. 北京：宇航出版社，1993.
[21] 王则力，巨亚堂，张　凯，等. 涡轮叶片辐射热冲击疲劳试验应力温度场模拟仿真[J]. 航天器环境工程，2019，36(4)：307—312.