文章编号:1671-4598(2018)02-0021-03 DOI:10.16526/j.cnki.11—4762/tp.2018.02.006 中图分类号:TP183 文献标识码:A

基于 PCA 和随机森林的故障趋势预测方法研究

王梓杰[1]，周新志[1,2]，宁 芊[1,2]
(1. 四川大学 电子信息学院，成都 610065；2. 电子信息控制重点实验室，成都 610036)

摘要：故障预测和健康管理技术（PHM）在现代工程系统中能够在系统具备较高复杂度的情况下，有效保障其可靠性和安全性；在机械故障诊断中对于采集到的原始数据的高维特征量的处理较为复杂，并且在实际应用中趋势预测的精度要求较高，针对该问题提出一种基于主成分分析（PCA）与随机森林算法的轴承故障趋势预测方法；该方法利用 PCA 对提取的原始轴承数据特征量进行线性降维，并选取其中主成分特征量，输出非线性时间序列数据；原始数据经过 PCA 处理得到非线性时间序列，将该序列作为随机森林算法的输入进行故障趋势预测，并把预测结果与 BP 神经网络模型预测的结果进行对比，结果表明随机森林在故障趋势预测上在精度相较于 BP 神经网络有显著提高，是一种有效的故障趋势预测方法。

关键词：趋势预测；PCA；故障诊断；随机森林；PHM

Study on the Fault Trend Prediction Method Based on PCA and Random Forest

Wang Zijie[1]，Zhou Xinzhi[1,2]，Ning Qian[1,2]
(1. College of Electronic and Information Engineering，SichuanUniversity，Chengdu 610065，China；
2. Science and Technology on Electronic Information Control Laboratory，Chengdu 610036，China)

Abstract：Fault prediction and health management system (PHM) can effectively guarantee the reliability and safety of modern engineering system under the condition of high complexity. In mechanical fault diagnosis with high dimensional characteristic quantity of raw data collected for the more complex，and in the practical application trend prediction precision，this paper presents a method based on principal component analysis (PCA) method to predict bearing fault trend and random forests algorithm. The method uses PCA to reduce the characteristic data of the original bearing data，and selects the principal component characteristic quantity to output the nonlinear time series data. The original data are processed by PCA nonlinear time series，the sequence as a random forest algorithm input fault trend prediction，and the prediction results were compared with the BP neural network model prediction results，results show that the random forest in the fault trend prediction in precision compared with the BP neural network is improved，is a method of forecasting an effective fault trend.

Keywords：trend prediction；PCA；fault diagnosis；random forest；PHM

0 引言

现代工业科技在信息化技术发展下，航天、通信和工业等各领域工程系统日趋庞大复杂，考虑到复杂系统的可靠性、安全性和经济性，以诊断与预测技术为核心的 PHM[1-2]（故障预测和健康管理系统）技术成为设备与系统保障的重要基础和技术支撑。PHM 主要包括故障诊断、故障预测和健康管理三个核心部分，其中故障诊断预测又可以分为故障分类[3]和趋势预测[4]等方向，目前的故障趋势预测主要通过传感器提取机械部件的时间序列物理量进行分析诊断，这些时间序列往往是非线性的，对于这类问题，常常用机器学习算法解决。文献［5］等基于神经网络信息融合对舵面系统故障趋势进行预测，但是神经网络在趋势预测中收敛速度缓慢[6-7]，同时网络的运算和结构参数依靠经验设置，调参优化缺乏理论指导；文献［8］等人使用 HMM/SVM 串联结构模型进行联合预测，取得优于任一单一算法的故障预测效果；文献［9］等人提出一种基于 ARMA 的趋势预测方法，但是容易出现调参复杂的问题。在实际的故障趋势预测中，往往具有多组物理量[10]，同时针对每一组时间序列的非线性数据，都可以提取很多频域和时域特征量用于趋势预测和故障分类[11]，而在将特征量输入算法作为趋势预测前，为了减少运算量提高精度，往往需要去除特征量中的冗余和干扰性的数据，这些数据无法准确反映趋势并且有重负数据冗余，因此在预测之前对数据进行降维预处理在某些应用场景下能显著提高预测精度，例如 PCA、KPCA 等特征降维与特征融合方法[12]。而随机森林算法[13]（Random forest）是利用多棵树对样本进行训练并预测的一种算法，它既可以应用在分类问题中，也可以用来做回归分析。随机森林相对于传统的决策树算法，具有不剪枝也能避免数据过拟合的特点，同时具备很快的训练速度，并且参数调整简单，在默认参数下往往就能够具备较好的回归预测效果。文中使用轴承退化过程的实验数据，选取 BP（back propagation）神经网络模型作为参照模型进行趋势回归效果比较。

1 特征提取与 PCA 降维处理

在机械轴承故障趋势预测中，由于环境噪声和设备的工况因素，传感器采集到的数据一般带有噪声，对这些时间序列物理量直接进行处理受噪声干扰较大得到的预测精度不高；在趋势预测中，机械的退化与故障反映在时序波形中有时并不能及

收稿日期：2017-10-27； **修回日期**：2017-11-30。

作者简介：王梓杰(1993-)，男，湖南株洲人，硕士研究生，主要从事模式识别方向的研究。

通讯作者：周新志(1966-)，男，四川成都市人，教授，硕士生导师，主要从事智能控制、信息与人工智能交叉等方向的研究。

时反映故障的开始时间，而是存在一定的时移；因此对传感器采集到的数据进行时域和频域的特征提取，本文所使用的数据集为，并且在不清楚不同特征量对于趋势预测的贡献率和相关度的情况下进行趋势预测往往得到的结果并不理想，因此在没有足够物理含义和先验知识的情况下，需要采取方法对特征量进行降维处理。

主成分分析[14]（Principal Component Analysis，后文简称为 PCA）是最常用的线性降维方法，对于原有的高维特征数据，利用坐标变换的思想，通过线性关系的投影，将高维的数据映射到低维的数据空间中表示，数据的对应关系并非简单的将原有高维数据进行信息量的删减，而是在高维向低维的坐标映射中对相关性特征量进行了整合，得到之前特征量的协方差矩阵，这里的特征量是一个经过重构的全新正交特征量。一方面去除原始数据中各维度数据间的线性关系对于最终分类或者预测算法的精度影响，另一方面，在样本数据不多，但是数据本身维度却相对较高的情况下提高算法分类或者预测的精度。得到低维度的特征量后，保留占据绝大多数影响的特征量，能在保留住较多的原数据点的特性的同时进一步降低特征数据的维度。PCA 的计算过程中不需要人为的设定参数或是根据任何经验模型对计算进行干预，最后的结果只与数据相关。但是，如果用户对观测对象有一定的先验知识，掌握了数据的一些特征，却无法通过参数化等方法对处理过程进行干预，可能会得不到预期的效果。是丢失原始数据信息最少的一种线性降维方式。因为 PCA 相对于其他的降维方法，对于原始数据的信息和关联性丢失较少。设定一个 PCA 的执行步骤如下：

1）构建 $m*n$ 阶的变量矩阵，其中 m 为样本数量，n 为原始数据的维数；

2）将 $m*n$ 阶的变量矩阵 $\boldsymbol{X}$ 的每一行，即原始数据的一个属性，进行数据的归一化处理；

3）求出协方差矩阵 $\boldsymbol{C}$，并对其特征值和特征向量进行求解；

4）将特征值从大到小进行排序，选择其中最大的 k 个，然后将其对应的 k 个特征向量分别作为列向量组成特征矩阵 $\boldsymbol{M}$；

5）即可以求得原 n 维的原始高维数据降维到 k 维后的数据 $\boldsymbol{Y}=\boldsymbol{XM}$。

矩阵 $\boldsymbol{Y}$ 是由数据协方差矩阵前 k 个最大的特征值对应的特征向量作为列向量构成的。这些特征向量形成一组正交基并且最好地保留了数据中的信息。

2 决策树与随机森林算法

2.1 决策树

相较于传统的神经网络和贝叶斯算法，决策树是以实例为基础的算法，通过不断的对样本归纳学习从而对分类以及预测等问题进行概率计算。决策树本身的构造并不需要相关样本数据领域的先验知识或者参数设置，因此，决策树很适用于探索性的应用。决策树本身是一个树结构（可以是二叉树或非二叉树）。它表示对象属性和对象值之间的一种映射，树中的每一个节点表示对象属性的判断条件，其分支表示符合节点条件的对象。树的叶子节点表示对象所属的预测结果。使用决策树进行决策的过程就是从根节点开始，测试待分类和待遇测项中相应的特征属性和特征值，并按照其值选择输出分支，将叶子节点存放的类别作为决策结果。构造决策树的关键步骤是分裂属性。所谓分裂属性就是在某个节点处按照某一特征属性的不同划分构造不同的分支，其目标是让一个分裂子集中待分类项属于同一类别。

在此基础上 J. Ross Quinlan 于 1986 年提出 ID3 算法，采用信息增益最大的特征；Breiman 等人于 1984 年提出 CART 算法利用基尼指数最小化准则进行特征选择；J. Ross Quinlan 于 1993 年提出 C4.5 算法，采用信息增益比选择特征。

2.2 随机森林

随机森林（Random Forest）是 Leo Breiman 和 Adele Cutler 在 2001 年提出的一个新的组合分类器算法，在此之后，Deitterich 在模型中引入了随即节点优化的思想，对随机森里进行了进一步完善，运用了 Leo Breiman 的“套袋”思想构建了控制方差的决策树集合。随机森林算法利用多个 CART（Classification And Regression Tree）作为元分类器，用套袋算法制造有差异的训练样本集，同时在构建单棵树时，随机地选择特征对内部节点进行属性分裂。因此随机森林能较好容忍噪声，并且具有较好的分类性能。实际应用中随机森林作为一种多功能的机器学习算法，除了执行回归、分类的任务，同时也用于处理缺失值、异常值以及其他数据探索中，作为一种降维手段。通常随机森林通过以下步骤运作：

1）我们设定一个样本个数为 N 的样本集，M 表示变量的数目；

2）每个节点都将随机选择 m（$m<M$）个特定的变量，然后运用这 m 个变量来确定最佳的分裂点。在决策树的生成过程中，m 的值是保持不变的；

3）从样本集（N 个样本）中以可放回取样的方式，取样 N 次，形成一组训练集（即 bootstrap 取样）。并使用这棵树预测剩余类别并评估其误差。

4）对于每一个节点，随机选择 m 个基于此点上的变量。根据这 m 个变量，计算其最佳的分裂点。

5）每棵决策树都最大可能地进行生长而不进行剪枝（Pruning），通过对所有的决策树进行加总来预测新的数据。

3 基于随机森林的故障趋势预测

3.1 实验数据

本次针对随机森林算法在轴承诊断中的应用，选择美国辛辛那提大学智能系统维护中心提供的滚动轴承全寿命周期加速轴承性能退化实验数据进行趋势预测实验。该数据为提取的加速度时间序列，采样的时间间隔是 10 min，采样频率是 20 kHz，采样点数为 20480 个，实验数据记录了从轴承完好到发生故障的全寿命周期过程，总共 984 条数据，本文截取其中后期从正常运行工况到具备退化趋势的一段数据进行实验，图 1 是轴承运行后期的第 700 条数据的振动信号幅值图。

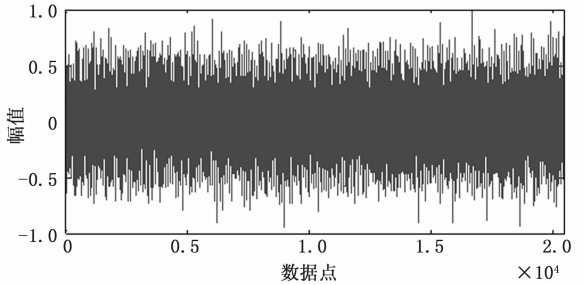

图 1 原始数据振动幅值图

3.2　特征提取与 PCA 降维

由于原始数据点数较多，且具有一定的噪声干扰，需要对原始数据进行压缩处理，提取特征量进行分析预测。参考文献(KPCA)，从每一节数据中提取 10 个频域特征量和 15 个时域特征量，共计 25 个特征量进行主成分分析，其中时域特征量如时域均值趋势如图 2，频域均方根值如图 3 所示。

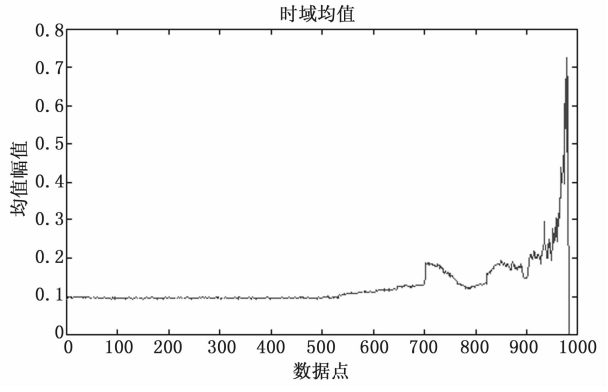

图 2　时域均值趋势

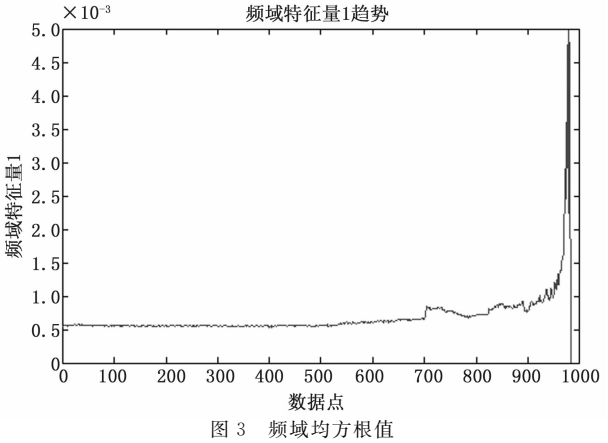

图 3　频域均方根值

对数据的趋势分析得到：从 500 点开始，数值呈现上升趋势，物理上的表现即反映轴承产生性能退化，并且在 700 点位置左右有第一个波峰。在所有 25 个特征值里，反映轴承实际退化趋势的有 18 个，为了降低数据冗余，提高预测精度，选取了这 18 个特征量进行 PCA 主成分分析对高维特征量进行降维，经过主成分分析得到前四个分量的贡献率如表 1 所示，其中分量 1 的贡献率超过 95%，为 96.3334%，依照 PCA 中选取贡献率位 85%以上的特征分量的原则，选择贡献率最高的分量作为随机森林预测效果的实验数据。

表 1　部分特征分量贡献率　　%

特征分量贡献情况				
特征分量	分量 1	分量 2	分量 3	分量 4
贡献率	96.3334	0.6281	0.0213	0.0163

3.3　实验方案及结果分析

3.3.1　随机森林预测模型构建

根据所采用的实验数据和随机森林的输入输出和结构，首先确定训练集和预测数据，参考数据分析结果，将 PCA 降维处理后得到的 984 个数据点中能正确反映轴承故障退化趋势的数据段中，701～900 数据点作为训练集，901～920 数据点作为预测数据，并建立训练集的训练样本特征空间 $S=[\boldsymbol{X},\boldsymbol{Y}]$，其中 $\boldsymbol{X}$ 为训练集样本空间如下：

$$\boldsymbol{X}=\begin{bmatrix} x_1 & x_2 & \cdots & x_{25} \\ x_2 & x_3 & & x_{26} \\ \vdots & \vdots & \cdots & \cdots \\ x_{185} & x_{186} & \cdots & x_{199} \end{bmatrix} \tag{1}$$

$$\boldsymbol{Y}=\begin{bmatrix} x_{26} \\ x_{27} \\ \vdots \\ x_{200} \end{bmatrix} \tag{2}$$

$\boldsymbol{X}$ 的列数为 26，为预测的步长，试验中分别选择 10、15、20、25 和 30 作为步长，实验结果显示当步长为 25 时随机森林预测模型具备最佳的预测效果，因此预测步长为 25。随机森林的树的数量选定 100～1000，以 100 为步长步进，得到的结果为树的数量设定为 500 时具有较好的预测精度。mtry 设置为 25，其他参数设置为默认值。

3.3.2　实验结果分析

为了验证本文采用的随机森林的预测效果，选取 BP 神经网络对数据进行预测比较两者的预测精度。选用 R 方和 RMSE 以及 MSE 作为衡量预测值和实际值拟合优度的标准，图 4 为原始数据点、随机森林预测数据点和 BP 神经网络预测数据点对比图。

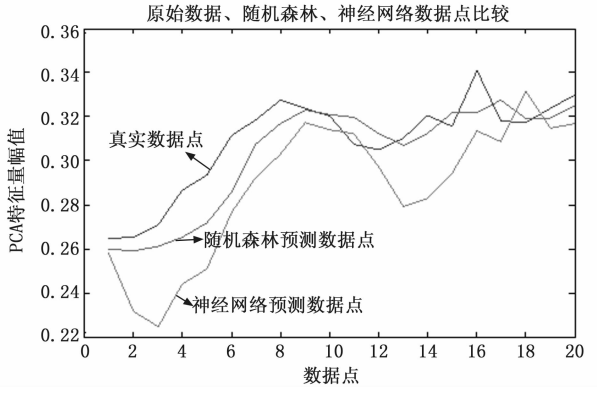

图 4　随机森林与 BP 神经网络对比图

从图 4 可以看到，神经网络在较为平缓的部分预测值就出现了较大的偏差，并且有明显的预测延迟的情况，而随机森林的预测趋势不但在较为平缓的地方和实际值一致，并且很好的反映了真实值在出现较大波峰时的趋势情况，不仅实际反映退化趋势，同时具备精度较高的预测数值。表 2 为随机森林算法和 BP 神经网络算法预测效果的 *RMSE* 值、*R* 方值以及 *MSE* 值的比较结果。可以看到随机森林模型的 *R* 方值为 0.9257，相比 BP 神经网络模型的 0.8077 提高了 14.6%；*RMSE* 值相对于神经网络，降低了 55%；随机森林模型的 *MSEMSE* 值相较于 BP 神经网络的 *MSE* 值要小一个数量级。

表 2　算法预测结果参数比较

随机森林与 BP 神经网络预测结果参数比较			
参数	*R* 方	*RMSE*	*MSE*
随机森林	0.9257	0.0119	0.000007
BP 神经网络	0.8577	0.0262	0.000058

(下转第 26 页)