# 工业物联网中涡轮风扇发动机的预测性维护研究

**作者**：黄锐，李承幸，黄旺辉，王彬，冯永

**单位**：重庆大学，计算机学院

**案例版权**：该案例归重庆大学计算机学院所有

**涉及的知识点**：工业物联网预测性维护、数据预处理、PCA降维、线性回归、决策树回归、支持向量机、深度神经网络、逻辑斯蒂回归

**案例来源及案例真实性情况**：该案例来源于重庆大学计算机学院专业硕士（电子信息）课程《大数据架构与技术》中的学生精选汇编课程设计。

**摘要** 工业物联网将具有感知、监控能力的各类采集、控制传感器或控制器，以及移动通信、智能分析等技术不断融入到工业生产过程各个环节。这种方式大幅度提高了制造效率，改善产品质量，降低产品成本和资源消耗，最终实现将传统工业提升到智能化的新阶段。本案例选取工业物联网中的预测性维护问题作为主题，使用来源于美国宇航局网站的涡轮风扇降级模拟数据集作为实验数据，结合数据挖掘、机器学习的相关方法进行涡轮风扇发动机的预测性维护研究。本案例核心内容有：（1）对原始涡轮风扇降级模拟数据集进行分析与可视化，并使用PCA降维等方式进行数据预处理；（2）使用线性回归、广义线性回归、决策树回归三种方法进行无论风扇发动机的剩余使用寿命预测；（3）使用支持向量机、深度神经网络、逻辑斯蒂回归三种方法进行涡轮风扇发动机的工作状态判定。

**关键词**：工业物联网，预测性维护，数据挖掘，机器学习

## 1 引言

该教学案例来源于重庆大学计算机学院专业硕士（电子信息）课程《大数据架构与技术》中的学生精选汇编课程设计。该案例的关键问题为工业物联网中的预测性维护，以涡轮风扇发动机的剩余使用寿命预测与运行状态预测作为具体问题，需引导学生进行的主要内容有：（1）对原始涡轮风扇降级模拟数据集进行分析与可视化，使用适当方法进行数据预处理；（2）选取适当的数据挖掘或机器学习方法对涡轮风扇发动机的剩余使用寿命进行预测；（3）选取适当的数据挖掘或机器学习方法对涡轮风扇发动机的工作状态进行判定。

## 2 背景介绍

工业物联网是将具有感知、监控能力的各类采集、控制传感器或控制器，以及移动通信、智能分析等技术不断融入到工业生产过程各个环节，从而大幅提高制造效率，改善产品质量，降低产品成本和资源消耗，最终实现将传统工业提升到智能化的新阶段。从应用形式上，工业物联网的应用具有实时性、自动化、嵌入式(软件)、安全性、和信息互通互联性等特点。

中国工业物联网产业链还处在形成初期，产业链条的界定和分工还不完全明晰，但随着产业整体竞争力的快速提升，行业处于爆发前期。再加上全国工业物联网数据创新应用大赛等赛事连连举行，都强调了工业界大数据应用的重要程度，也反映出工业界对大数据智能化人才的迫切需求。因此一定比例的工业物联网案例对于专业硕士的培养是必要的。在本案例中，我们选取工业物联网中的预测性维护作为主要研究问题，借助大数据分析与处理、数据挖掘、机器学习的相关方法，以满足培养具备创新能力与实践能力的人才的需求。

## 3 内容

该案例的主要内容主要分为三个小节，分别为数据集分析与预处理、涡轮风扇发动机剩余使用寿命预测，以及涡轮风扇发动机工作状态判定。

3.1 数据集分析与预处理

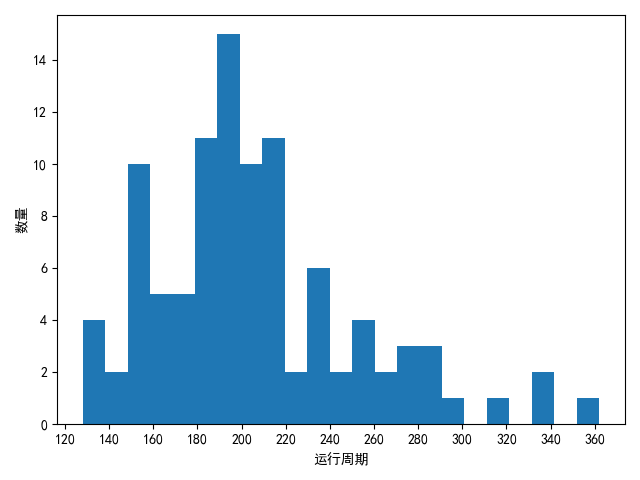
本案例采用的数据集为涡轮风扇降级模拟数据集，可从美国宇航局网站获取：https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository/。

本数据集共包含4组数据，每一组包含1个训练数据文件，1个测试数据文件和1个真实剩余使用寿命文件。以训练数据中一组数据为例，其包含了100台同型号的涡扇发动机的数据，共约两万多条数据。数据维度为26，包含1个发动机id，1个已运行周期数量，3个操作设置及21个传感器的读数。具体字段含义如下表所示：

|  |  |
| --- | --- |
| **字段名** | **含义** |
| unit number | 涡扇发动机编号 |
| time, in cycles | 使用时间，以周期为单位 |
| operational setting 1 | 操作设置1 |
| operational setting 2 | 操作设置2 |
| operational setting 3 | 操作设置3 |
| sensor measurement 1 | 1号传感器读数 |
| sensor measurement 2 | 2号传感器读数 |
| ... | … |
| sensor measurement 21 | 21号传感器读数 |

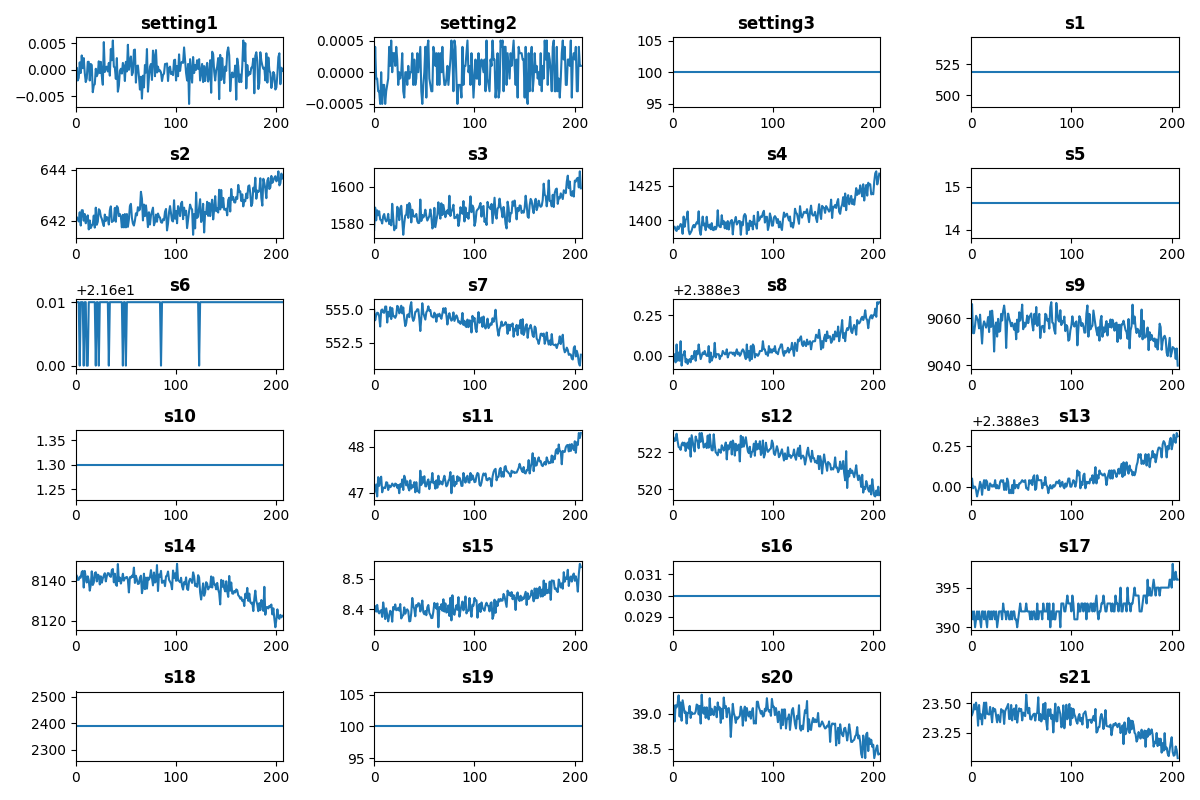
首先我们对数据集进行可视化展示以分析数据特征。这里给出可视化分析的两个例子：

（1）提取100台发动机的使用寿命，并以10个周期为基本单位绘制直方图，分析其大致分布，寿命分布直方图如下：



可以发现，此种型号的涡扇发动机寿命数据分布类似于正态分布，主要集中150-240个运行周期之间。

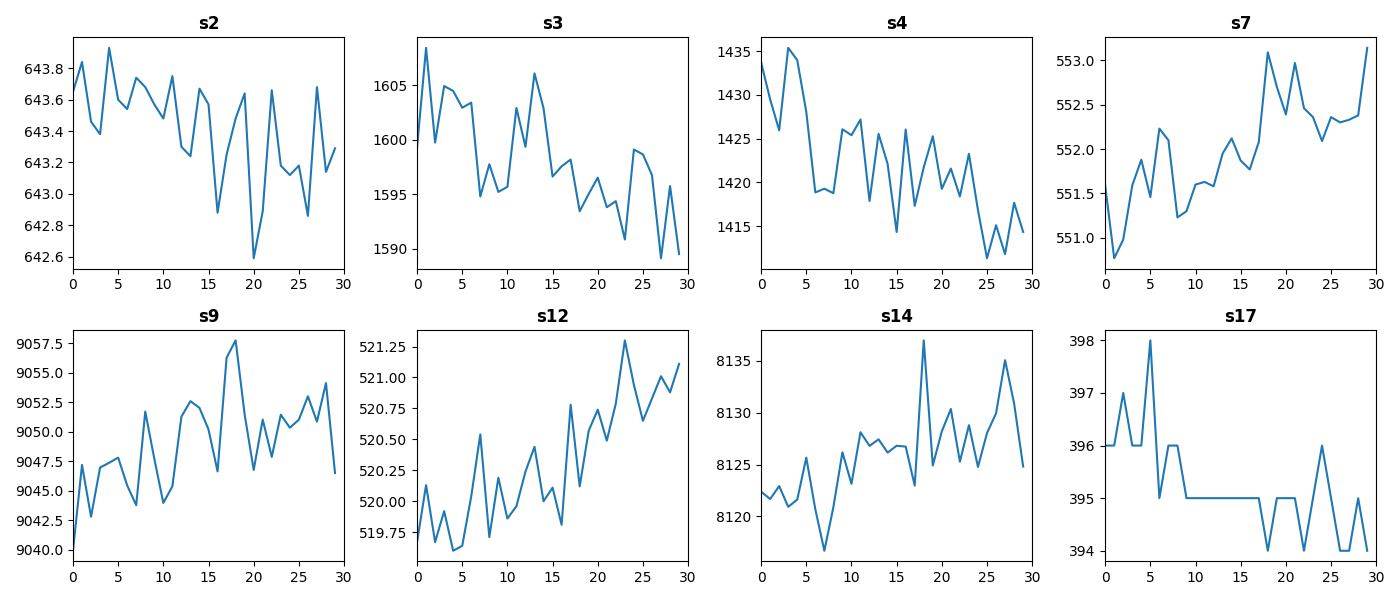
（2）绘制各数据字段随时间变化的折线图，以评估各项特征对于构建预测模型的作用。以1号发动机为例，其特征变化趋势图如下所示：



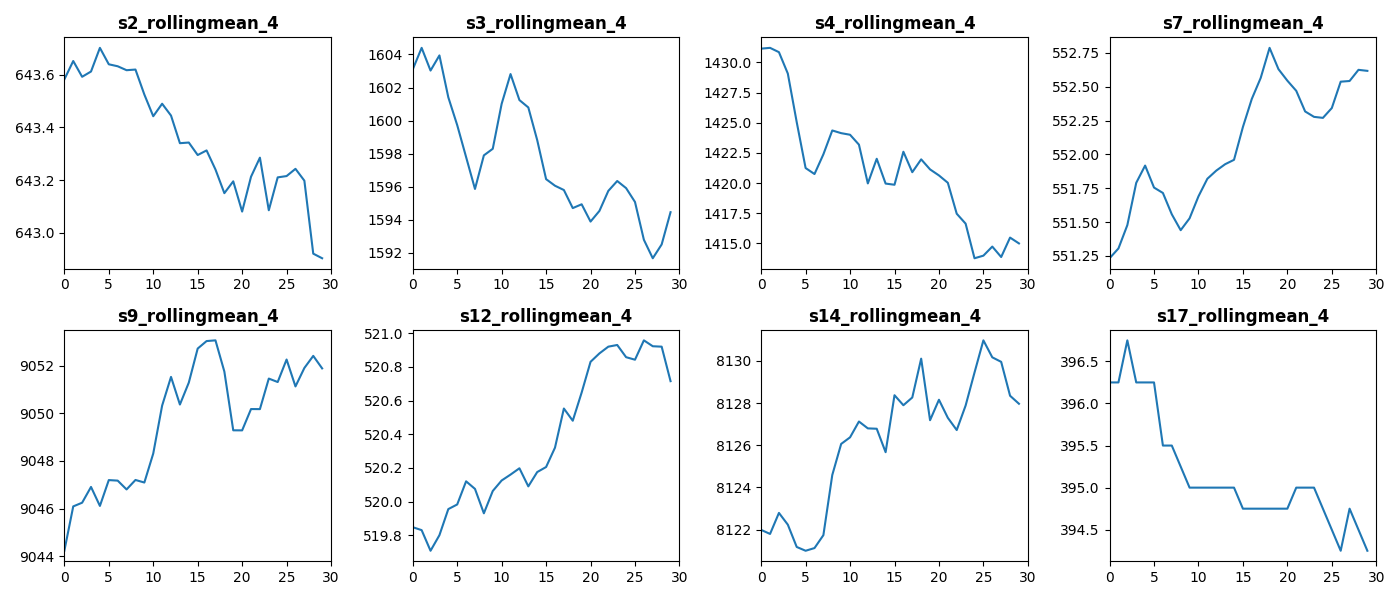
可以直观地看出，有部分特征没有随时间呈现规律性的变化，这种类型特征在构建预测模型中基本无用。因此，可以考虑对训练集和测试集均删除低方差特征。

数据集的预处理是数据挖掘中极其重要的步骤之一，有效的数据预处理往往能显著增强预测模型的性能。这里给出数据集预处理的三个参考步骤：

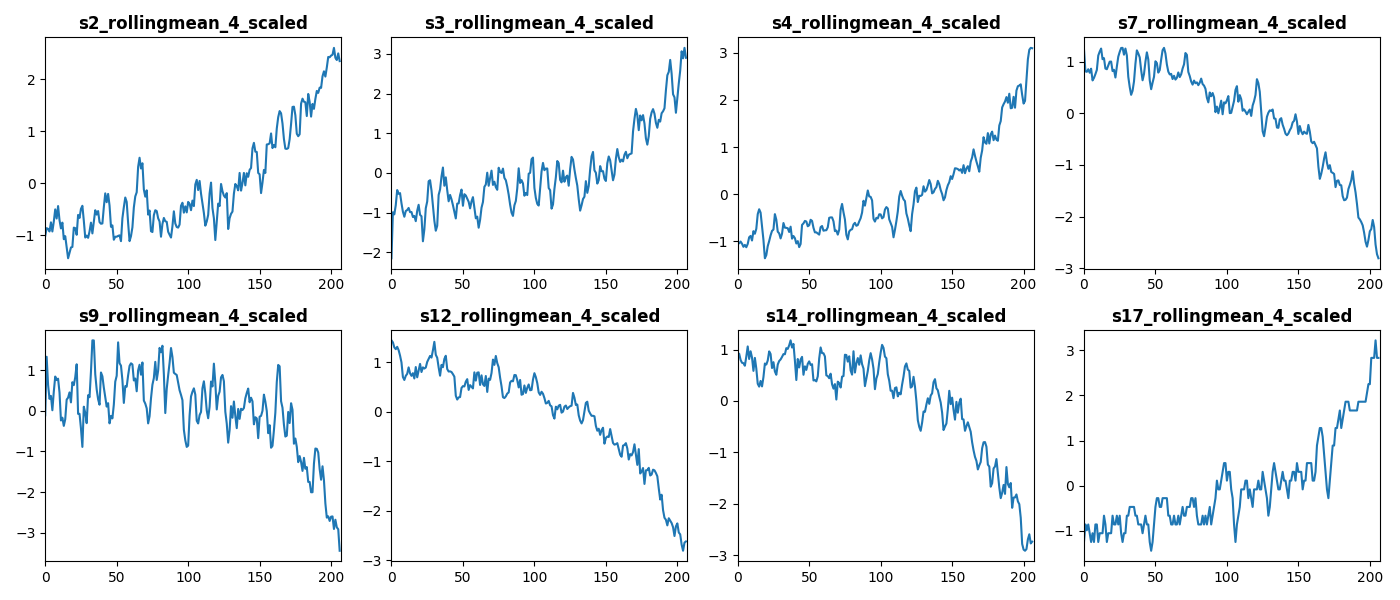
（1）对训练集和测试集删除低方差特征后，可得到8个共有特征。可视化展示如下。



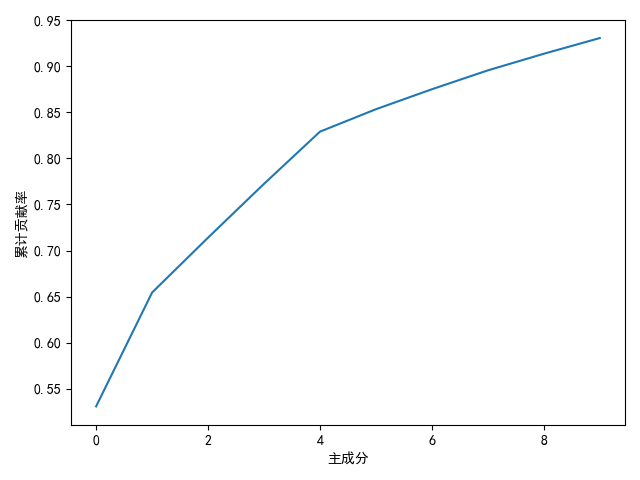
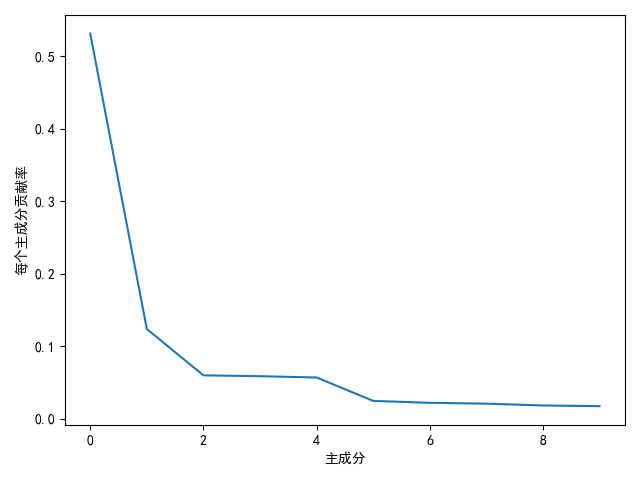
由于传感器读数会出现误差及产生不规则数据读数，可以使用移动平均法消除噪声。消除噪声之后的特征值展示如下。



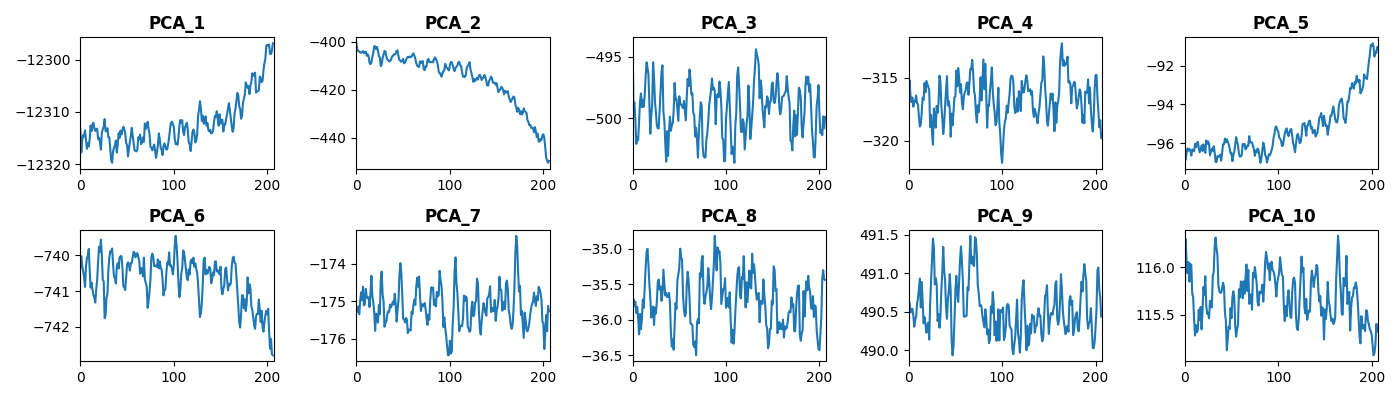
（2）对数据进行归一化与标准化，使每个特征的均值为0，标准差为1，可提升机器学习模型的性能。处理后的数据如下所示。



（3）由于在之前的步骤之中生成了许多的新特征，数据维度已大为扩展，因此需要降低数据维度，便于调节机器学习算法的参数，并优化预测解决方案。此处采用PCA进行降维，得出前10个主成分的单个贡献率和累计贡献率，由下图可知，前10个主成分的累计贡献率已达到90%。



再次进行全局归一化和标准化，生成最后的PC特征。PC特征如下图所示：



3.2 涡轮风扇发动机剩余使用寿命预测

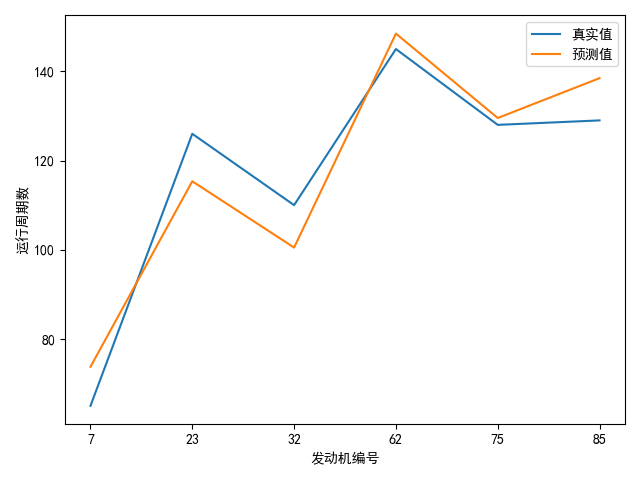
在本小节，我们将整个训练数据文件用作训练集，分别训练线性回归、广义线性回归及决策回归树模型。同时，我们还在测试数据中随机选取5台发动机的数据作为测试集，以均方根误差（RMSE）作为评价指标，从而选择最佳模型及参数。

三种方法通过调节多种参数组合，得到实验结果如下：

|  |  |  |
| --- | --- | --- |
| **模型** | **参数** | **均方根误差（RMSE）** |
| 线性回归 |  | 85.85 |
| 广义线性回归 | Lambda = 0 | 17.58 |
| 广义线性回归 | Lambda = 0.5 | 35.31 |
| 广义线性回归 | Lambda = 1 | 53.13 |
| 广义线性回归 | Lambda = 10 | 71.19 |
| 广义线性回归 | Lambda = 30 | 89.23 |
| 广义线性回归 | Lambda = 50 | 107.23 |
| 决策树回归 | maxDepth = 5, maxBins = 4 | 18.12 |
| 决策树回归 | maxDepth = 5, maxBins = 8 | 36.20 |
| 决策树回归 | maxDepth = 10, maxBins = 4 | 54.50 |
| 决策树回归 | maxDepth = 10, maxBins = 6 | 73.06 |

从结果可以发现，使用参数Lambda=0的广义线性回归获得最优的效果。进一步地，将广义线性回归用于整个测试集，并评估预测性能如下所示：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **发动机编号** | **7** | **23** | **32** | **62** | **75** | **85** |
| 真实值 | 65 | 126 | 110 | 145 | 128 | 129 |
| 预测值 | 73.75 | 115.35 | 100.49 | 148.46 | 129.54 | 138.47 |
| 差值 | -8.75 | 10.65 | 9.51 | 3.46 | 1.54 | 9.47 |



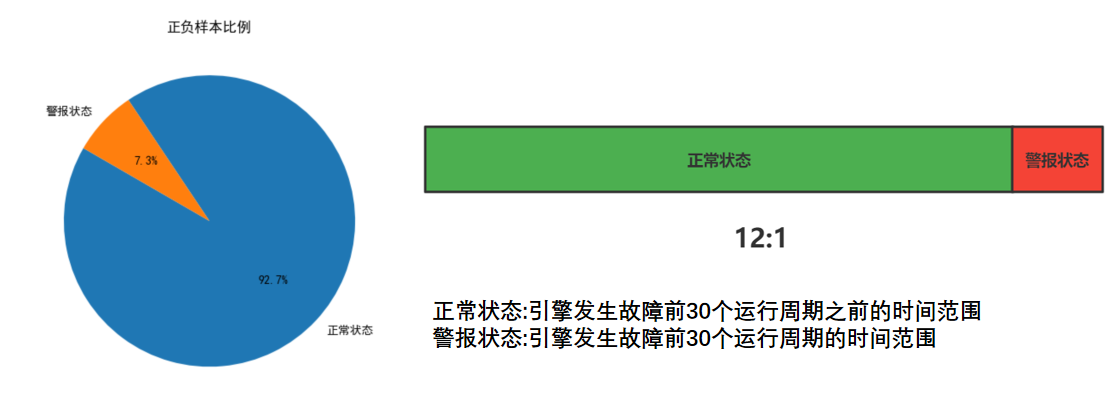
广义线性回归所得预测结果的部分展示如上图所示，预测均方根误差为19.04。虽然模型取得不错的效果，但还有进一步优化的空间。在本案例中，可以考虑进一步地调整模型超参数，或运用集成学习的方法，训练出更好的预测模型。

3.3 涡轮风扇发动机工作状态判定

涡轮风扇引擎在整个运行周期可以大致分为两种状态：正常状态和警报状态。正常状态指引擎出现问题时的前30个运行周期之前所处的状态，警报状态则是指离引擎出现问题时的最近30个周期。在本小节中，需要根据预处理之后的数据，对某一时刻的涡轮风扇所处的状态进行判定。该小节主要分为两个步骤：

（1）正负样本均衡

在原始训练数据集中，正常状态与警报状态的数据比例约为12:1，存在正负数据不均衡的情况，如下图所示：

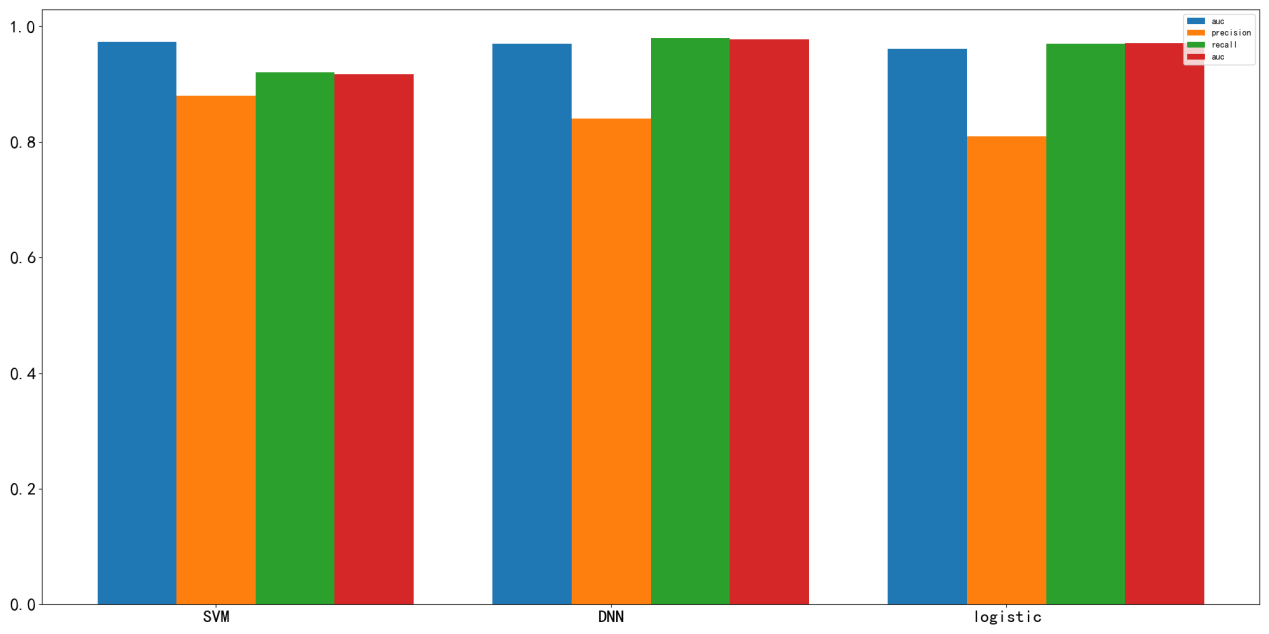


我们采用SMOTE算法对训练集中的正负样本进行平衡。SMOTE即合成少数类过采样技术。它是基于随机过采样算法的一种改进方案，由于随机过采样采取简单复制样本的策略来增加少数类样本，容易产生模型过拟合的问题，即使得模型学习到的信息过于特别而不够泛化。SMOTE算法的基本思想是对少数类样本进行分析并根据少数类样本人工合成新样本添加到数据集中。经过SMOTE算法进行正负样本均衡后，其正负样本比例为1:1。

（2）训练工作状态判定器

对涡轮风扇发动机的工作状态判定本质上为一个分类问题。我们分别使用支持向量机（SVM），深度神经网络（DNN）和逻辑斯蒂（logistic）回归的方法对引擎状态进行判定。最后模型采用准确率，精确率，召回率和auc值进行综合评价，选取表现最佳的分类器。

三种分类器的具体表现如下图所示。



综合四种指标考虑，DNN分类器在所有指标上表现最为均衡，为最佳模型。

## 4 小结

本案例选取工业物联网中的预测性维护问题作为主题，使用来源于美国宇航局网站的涡轮风扇降级模拟数据集作为实验数据，结合数据挖掘、机器学习的相关方法进行涡轮风扇发动机的预测性维护研究。主要内容包括原始数据集的可视化分析与处理、发动机使用寿命预测与工作状态判定。该案例主题新颖，结合了工业界现实需求与大数据分析与挖掘的多种理论与技术，可以充分增强学生的实践能力与理论基础。另外，本案例的内容仅为指导性的过程，在实际教学中，可保持基本研究内容不变，鼓励学生引入其它的数据预处理、数据挖掘、机器学习方法完成任务。

## 附录

1. 本案例提供配套的PPT、视频、数据集与代码等，发布于Github，链接为：https://github.com/Wanghui-Huang/CQU\_bigdata。

2. 本案例涉及到数据预处理以及多种机器学习算法，建议使用python语言进行编写，推荐的工具包有pandas（数据读取与预处理库），scikit-learn（机器学习算法库），matplotlib（可视化绘图库）。

3. 本案例参考文献如下：

[1] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.

参考内容：数据预处理理论，线性回归、广义线性回归、决策树、SVM、logistic回归算法

[2] Goodfellow I, Bengio Y, Courville A. Deep learning[M]. MIT press, 2016.

参考内容：深度神经网络DNN模型