设备疲劳断裂检测的数据采集分析优化方法

黄 啸[1] 叶序彬[2]
HUANG Xiao YE Xubin

摘 要 针对机械零部件疲劳断裂在数据采集与分析阶段存在的数据分析优化问题，提出了一种基于加权模糊 C 均值 - 对数变换 PCA 算法，通过在余弦相似度的基础上，引入了基于余弦值的欧几里得距离加权处理方法来度量数据的相似性特征，从而改进了传统模糊 C 均值算法，提高了对数据点的精准判断以及噪声检测率。然后针对降噪后的数据依然存在数据维度过高且难以分析数据特性等问题，在改进的加权模糊 C 均值算法基础上利用对数变换对 PCA 算法进行了改造，不仅降低了冗余数据噪声而且降低了数据维度，最后进行了真实样本数据集测试，达到预期。

关键词 疲劳断裂；数据采集；数据降噪；数据降维；PCA

doi：10.3969/j.issn.1672-9528.2025.03.036

0 引言

疲劳断裂是指机械设备零部件由于重复和循环载荷引起的弱化，导致裂纹的形成和扩展或者结构损伤，设备零件疲劳断裂是承载构件结构失效的最突出原因之一 [1-2]。设备零部件疲劳断裂的发生是最危险的事故之一，通常来说断裂之处都在被遮挡的零件局部，即使人为的检查都很难发现，况且车间等场所设备众多，仅靠人工去筛查往往费时费力。力学检测实验室作为一种重要的机械设备零部件疲劳断裂检测场所，通常在工程研究中被用于对机械或金属运动构件的疲劳断裂和生命周期的检测，覆盖低周疲劳、高周疲劳、旋转弯曲疲劳和断裂韧性等试验类型，因此对力学检测实验室的轴向试验机、旋弯疲劳试验机、裂纹扩展系统等设备和系统的数据采集及分析显得尤为重要。机械设备零件疲劳断裂的数据采集分析和过滤是实现设备性能分析的一个重要前提，若不能及时发现零部件疲劳断裂损伤，将对设备造成极大破坏，引发事故。

但是目前，在制造业领域力学检测方面，尤其对于机械设备零件疲劳断裂的检测数据的分析存在以下问题：由于力学检测实验场所环境复杂，机械和检测设备多，采集的数据量大、多源异构数据多，且离散分布，缺乏检测过程的数据校对和防错机制分析，检测数据质量分析不足，这就不可避免的存在噪声数据。目前传统主成分分析算法对于数据降维和去噪的应用比较广泛，然而对于力学检测实验场所的多源异构、离散型的数据分析和去噪没有一个成熟且适用的技术方案。

针对以上问题，引入了基于加权模糊 C 均值 - 对数变换 PCA 算法，实现了疲劳断裂检测非线性数据的去噪和降维的优化，更大地保留了原始数据信息。

1 数据采集架构

力学检测实验场所车间的数据采集有多种方式，然而最常见的主要有 3 种方式，即直接采集、间接采集和导入采集，如图 1 所示。

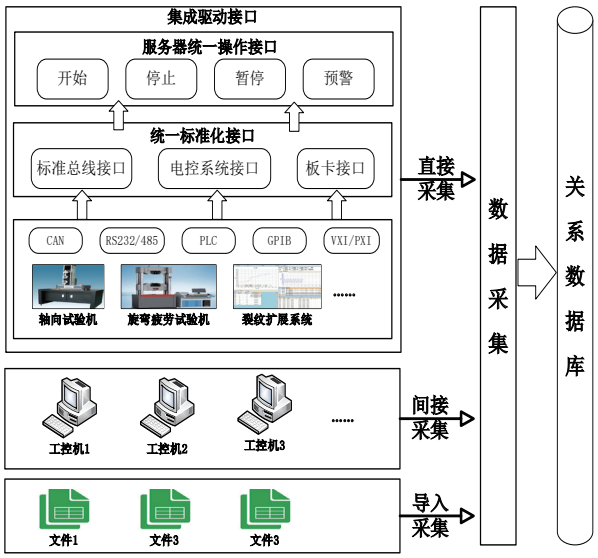

图 1 数据采集架构

直接采集是系统通过采集适配器直接采集设备的通信参数、疲劳断裂数据等。间接采集是通过采集工控机的网口、

1. 中国航空发动机集团有限公司 北京 100097
2. 中国航发北京航空材料研究院 北京 100097

GPIB 口、USB 口、串口等接口信息，获取设备产生的数据。导入采集是通过采集软件配置过滤当前需要解析的文件类型，通过 C++ 文件句柄打开文件，根据采集软件产生文件的格式解析文件，读取文件中的数据，最终上传至服务器分析优化。

2 主成分分析算法

2.1 算法概述

数据时代，每时每刻都会产生海量的信息，在这些信息之中有需要的，也有无用的、冗余的噪声数据，如何从这些海量的大数据中提取有用的信息是一个关键问题。简而言之，如何减少数据维度的同时还能基本保持数据的原始信息特征不改变，这就需要一种数据降维方法 - 主成分分析算法。

主成分分析算法（PCA）[3]，是一种在人工智能、经济金融、医疗医学、社会科学等领域广泛使用的线性数据压缩降维方法。将原来复杂的样本数据指标通过空间线性变换降维成少数几个带有大量样本信息的主成分，即从高维度降到低维度，在减少时间复杂度的同时，基本保留了原始样本数据的特征。

2.2 算法原理

假设有数据集：$X=\{x_1,x_2,\cdots,x_n\}$，将其降到 k 维，方法如下：

（1）对原始数据进行标准化处理，以消除不同变量对分析结果的影响，计算公式为：

$$\bar{x}=\frac{1}{x}\sum_{i=1}\boldsymbol{x}_i \tag{1}$$

（2）计算标准化后的数据协方差矩阵，样本 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的协方差计算公式为：

$$\begin{aligned}\boldsymbol{C}(\boldsymbol{X},\boldsymbol{Y})&=\boldsymbol{E}[(\boldsymbol{X}-\boldsymbol{E}(\boldsymbol{X}))(\boldsymbol{Y}-(\boldsymbol{Y}))]\\&=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)\end{aligned} \tag{2}$$

当样本是 n 维数据时，它们的协方差实际上是协方差矩阵（对称矩阵），计算公式为：

$$\boldsymbol{C}(x_1,x_2,\cdots,x_n)=\begin{bmatrix}\boldsymbol{C}(x_1,x_1) & \boldsymbol{C}(x_1,x_2) & \cdots & \boldsymbol{C}(x_1,x_n)\\ \boldsymbol{C}(x_2,x_1) & \boldsymbol{C}(x_2,x_2) & \cdots & \boldsymbol{C}(x_2,x_n)\\ & & \cdots & \\ \boldsymbol{C}(x_n,x_1) & \boldsymbol{C}(x_n,x_2) & \cdots & \boldsymbol{C}(x_n,x_n)\end{bmatrix} \tag{3}$$

（3）计算协方差矩阵 $\boldsymbol{C}$ 的特征值与特征向量。

（4）通过特征值大小将特征向量排列成 k 行矩阵。

（5）计算 $\boldsymbol{Y}=\boldsymbol{PX}$，即为降维后的数据集。

3 基于加权模糊 C 均值 - 对数变换 PCA 算法去噪降维优化

由于检测场所环境复杂，采集的数据具有数量大、异构多源且离散的特点，而且存在着空值、错误值或异常值，这类数据统称为噪声数据。噪声数据通常会严重影响疲劳断裂的检测效果，因此需要对采集的数据先进行噪声过滤再降维，以最大限度的消除一些噪声和冗余信息，从而提高数据的信噪比 [4]，使得检测达到理想效果。

3.1 算法改进原理

根据分析数据的特征，选取模糊 C 均值聚类算法 [5] 作为去噪过滤算法。通常来说数据降噪方法有多种方式，基于距离和相关性的两种方法最为常见。前者是基于欧几里得距离的噪声过滤方法 [6]，根据计算距离的阈值来断定噪声。后者是基于余弦相似度进行度量，余弦取值范围为 [-1,1]，其中 -1 为完全不相似，1 为完全相似，0 表示它们之间是独立的。根据 $\cos\theta$ 的特点，其 $\cos\theta$ 值与角度成反比，$\cos\theta$ 越大，角度越小，表示两个数据所携带的特征越相似。余弦相似度公式为：

$$S=(x,y)=\cos\theta=\frac{\vec{x}\vec{y}}{|x||y|}=\frac{\sum_{i=1}^{n}x_i y_i}{\sqrt{\sum_{i=1}^{n}(x_i)^2}\sqrt{\sum_{i=1}^{n}(y_i)^2}} \tag{4}$$

通过多次算法仿真实验结果得知，利用欧几里得距离的噪声过滤方法往往存在这样一个问题，面对复杂的场景采集的数据集大多是多源异构的，其中包含缺失值、空值、异常值等冗余的数据，且数据是非线性多方向的。欧几里得公式虽然计算简单，效率高，但是通常情况下只关注点对点的直线距离而忽略了数据的方向性特征，不同方向上的样本数据点经常聚为一簇，不能够差别的区分数据点，因此，为了弥补传统模糊 C 均值算法对这一问题的错误判断，在余弦相似度的基础上，提出了一种基于余弦值的欧几里得距离加权处理方法来度量数据的相似性特征，提高了对数据点的精准判断问题。

设 m_j 是样本数据集 j 簇中心，那么对此簇内的任意数据点 $x_t^{(j)}$ 加权处理，得到欧几里得距离加权公式：

$$d_m=(x_i,y_j)=\sin\left(x_t^{(j)},y_j\right)\cdot\sqrt{\left(x_t^{(j)}-y_j\right)^t\left(x_t^{(j)}+y_j\right)} \tag{5}$$

式中：$t=|y_j|$ 表示 j 簇内的样本数目；$x_t^{(j)}$ 表示 y_j 所在簇的所有样本点。

通过改进的传统模糊 C 均值算法较好的解决了样本数据集的噪声过滤问题，即使降噪后的数据集没有了噪声的影响，然而对于采集的实验车间多场景复杂的设备零部件疲劳断裂多源异构数据，依然存在数据维度过高且难以分析数据特性等问题。

从采集的数据特征来看，原始数据呈现非线性关系，如果还是按照传统的主成分分析方法对这些数据进行降维，那么会直接导致数据降维效果大大降低，第一主成分不能够包含众多的原始数据信息，使得检测结果偏差较大。因此，从

力学检测实验场所采集的多源异构、离散型的数据分析，做了大量函数测试，最终选择了对数变换函数对主成分分析算法进行数据降维。

设 x_{ij} 是原始数据集，令：

$$y_{ij}=\ln x_{ij}-\frac{1}{p}\sum_{t=1}^{p}\ln x_{it} \tag{6}$$

然后计算对数变换后的标准化样本协方差矩阵 $\boldsymbol{S}=(s_{ij})_{p\times p}$，其中：

$$s_{ij}=\frac{1}{n}\sum_{t=1}^{n}(y_{yi}-y_i)(y_{it}-\bar{y}_j) \tag{7}$$

$$\bar{y}_i=\frac{1}{n}\sum_{t=1}^{n}y_{it} \tag{8}$$

$$\bar{y}_i=\frac{1}{n}\sum_{t=1}^{n}y_{ij} \tag{9}$$

接着根据 $\boldsymbol{S}$ 计算样本的主成分。设 $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p$ 是 $\boldsymbol{S}$ 的 p 个特征根，$\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_p$ 是相应的标准化单位特征向量，则第 i 个主成分为 $\boldsymbol{F}_i=\sum_{j=1}^{p}a_{ij}\ln x_{ij}$。

3.2 改进算法的数据分析优化流程

基于加权模糊 C 均值 - 对数变换 PCA 算法去噪降维优化具体实现如下：

假设有非线性样本数据集 x_{ij}，首先对数据进行去噪处理，以提高数据质量。输入样本数据集 x_{ij} 和聚类中心样本数 t，然后带入加权模糊 C 均值算法模型，计算加权的欧几里得距离 d 和阈值 r，即利用欧几里得距离设置阈值 r，r 可选取为所有样本点到聚类中心的加权欧氏距离的均值，即：

$$r=\frac{\sum_{1}^{k}\sum_{1}^{t}d_m\left(x_t^{(j)},y_j\right)}{n} \tag{10}$$

比较 d 和 r 的大小，若 $d>r$ 时，则为噪声数据，将其删除，反之 $d<r$ 时，则留下该条数据，即降噪后的数据集。

然后对其构造 $y_{ij}=\ln x_{ij}\frac{1}{p}\sum_{t=1}^{p}\ln x_{it}$ 对数变换函数，求每个 x_{ij} 均值数据，利用原始数据集减去均值数据得到去中心化数据集，计算协方差矩阵，求其特征值与特征向量，最后对特征向量进行空间投影，即得到降维数据集，如图 2 所示。

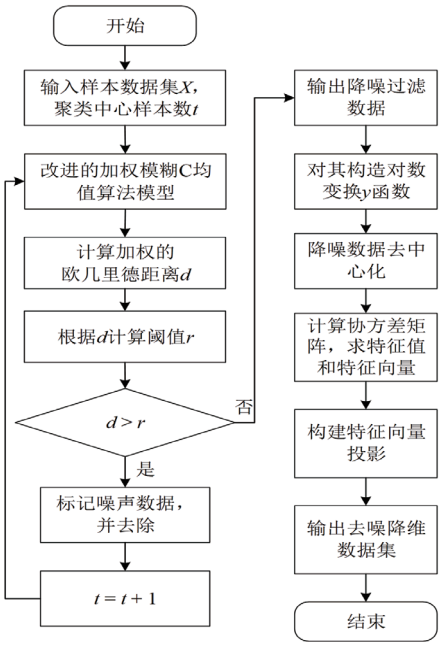

图 2 改进算法的数据分析优化流程

3.3 测试分析

针对改进的基于加权模糊 C 均值 - 对数变换 PCA 算法进行测试，以力学检测实验场所采集的多源异构、离散型数据集作为测试集进行测试，测试集中含有数据 10 000 条，其中缺失值 378 条，空值 121 条，异常值 239 条，共计 738 条，噪声数据占比为 7.38% 。算法处理的结果受聚类中心数样本数 t 的影响较大，在其他参数不变下，从 t=100 开始逐渐增大 t 的取值计算噪声检测率，取前 10 个主成分计算特征根和贡献率。将传统模糊 C 均值算法、传统主成分分析算法、改进的算法的处理结果进行比较，如表 1 所示。

从表 1 可以看出，当 $t<300$ 时，各个算法检测率基本

表 1 测试结果对比表

	传统模糊 C 均值算法		传统主成分分析		改进的算法			
t	噪声检测率 /%	时间 /s	特征根	贡献率 /%	噪声检测率 /%	时间 /s	特征根	贡献率 /%
100	84.32	0.31	5.675	88.134	84.76	0.32	5.675	94.234
200	84.41	0.35	1.428	8.321	84.98	0.36	1.428	6.543
300	84.57	0.42	0.123	7.532	85.21	0.44	0.123	5.755
400	84.85	0.51	0.098	6.875	85.73	0.56	0.098	4.876
500	85.87	0.64	0.075	5.467	86.41	0.69	0.075	3.536
600	86.02	0.78	0.045	4.289	87.43	0.83	0.045	2.846
700	86.88	0.97	0.022	3.980	88.97	1.05	0.022	1.913
800	87.54	1.31	0.008	2.390	90.51	1.37	0.008	0.645
900	89.65	1.57	0.005	1.461	93.15	1.67	0.005	0.324
1 000	91.21	2.15	0.002	0.537	95.86	2.21	0.002	0.241

注：噪声检测率 = 检测出的噪声数 ÷ 噪声总数；贡献率 = 对应的特征值 ÷ 所有特征值之和

（下转第 157 页）

参考文献：

[1] 刘长红，曾胜，张斌，等．基于语义关系图的跨模态张量融合网络的图像文本检索 [J]. 计算机应用 ,2022,42(10):3018-3024.

[2]ZHENG K F, WANG N, LIU J W, et al. An efficient multikeyword fuzzy ciphertext retrieval scheme based on distributed transmission for internet of things[EB/OL]. (2022-04-08) [2024-02-19]. https://doi.org/10.1002/int.22886.

[3] 杨帆，宁博，李怀清，等．基于语义增强特征融合的多模态图像检索模型 [J]. 浙江大学学报 (工学版), 2023, 57(2): 252-258.

[4] HUSSAIN S, ZIA M A , ARSHAD W. Additive deep feature optimization for semantic image retrieval[J].Expert systems with applications, 2021, 170(5):114545.

[5] 郜帅，侯心迪，刘宁春，等．多模态网络环境异构标识空间管控架构研究 [J]. 通信学报 ,2022,43(4):26-35.

[6] 赵鹏，马泰宇，李毅，等．融合全模态自编码器和生成对抗机制的跨模态检索 [J]. 计算机辅助设计与图形学学报，2021, 33(10): 1486-1494.

[7] 赵磊．基于深度学习的多模态数据特征提取与选择方法研究 [D]. 天津：天津大学，2015.

[8] 王晓莉．基于差分进化算法的思政多模态语料库智能构建 [J]. 微型电脑应用 ,2022,38(5):149-151.

[9] 尤博，彭开香．基于有效分类的多模态过程故障检测及应用 [J]. 上海应用技术学院学报（自然科学版），2015, 15(3): 242-247.

[10] 熊回香，杨滋荣，蒋武轩．跨媒体知识图谱构建中多模态数据语义相关性研究 [J]. 情报理论与实践，2019, 42(2): 13-18.

【作者简介】

周春良（1996—），男，河南开封人，硕士，助教，研究方向：光学、图像识别。

杨畅畅（1996—），男，河南商丘人，硕士，助教，研究方向：机器视觉、智能生产。

李晓辉（1995—），男，河南平顶山人，本科，中级工程师，研究方向：计算机视觉、边缘计算。

（收稿日期：2024-11-15）

（上接第 152 页）

一致，当 $t>300$ 时，本文改进的算法逐渐体现出优势，时间稍微增大的同时噪声检测率逐渐提高。而且，经过改进的算法处理的新数据集得到的第一主成分包含的信息比传统算法承载更多的信息量，因此可用较少的主成分提取更多的原始信息，降维效果更显著，改进的算法模型更具实用性。

4 结语

通过对机械零部件疲劳断裂数据采集及分析存在的问题进行梳理，提出了一种基于加权模糊 C 均值 - 对数变换 PCA 算法用于疲劳断裂检测数据的分析优化处理。与传统的算法相比，不仅能够提高噪声检测率，而且第一主成分包含更多原始信息，提高了数据降维效果，不仅加强了试验数据的安全管理，而且提高了疲劳断裂测试人员的管理能力。

参考文献：

[1] 王利东．机械工程中金属材料的疲劳与断裂分析 [J]. 产品可靠性报告，2024(6):117-118.

[2] AMOOIE M A, LIJESH K P, MAHMOUDI A, et al. On the characteristics of fatigue fracture with rapid frequency change[J]. Entropy, 2023, 25(6):840.

[3] 郭尚志，廖晓峰，李刚，等．基于PCA的大数据降维应用[J]. 计算机仿真，2024, 41(5):483-486.

[4] 孙家祥，胡春玲．方差参数和信噪比参数特定于父节点的全局耦合模型 [J]. 电脑知识与技术，2023, 19(27): 9-12.

[5] CHEN T, KUO D, CHEN C Y Z. Retraction note: fuzzy c-means robust algorithm for nonlinear systems[J].Soft computing，2023, 3(28): 2769-2775.

[6] 汪杨海，贺细平．扩展欧几里德算法改进探讨 [J]. 电脑与信息技术，2018, 26(6):12-14.

【作者简介】

黄啸（1986—），通信作者（email:weco_x@163.com），男，浙江宁波人，硕士，高级工程师，研究方向：材料力学性能表征。

叶序彬（1980—），男，湖北大冶人，硕士，高级工程师，研究方向：材料力学性能表征。

（收稿日期：2024-11-17）