文章编号：1006-1355(2014)05-0150-06

基于LLTSA的转子故障数据集降维方法

袁德强[1,2]，赵荣珍[1,2]

（1. 兰州理工大学 数字制造技术与应用省部共建教育部重点实验室，兰州 730050；
2. 兰州理工大学 机电工程学院，兰州 730050）

摘 要：特征数据集降维是机械故障智能诊断的关键步骤之一。常见的数据集降维方法难以准确的从高维非线性数据集中获取反映转子运行状态的敏感特征信息，导致故障模式识别精度降低。局部切空间排列算法（LTSA）对部分高维非线性数据集达到较好降维效果。但该算法不适合处理高曲率分布、稀疏不均匀分布等高维数据源。为此，在LTSA算法的基础上结合线性分块思想，提出线性局部切空间排列算法（LLTSA）。该算法充分考虑了数据集的整体与局部结构，将数据样本空间划分为一组最大线性块，使降维后的同类数据具有更好的聚类性。通过高维非线性转子振动数据时域特征数据集对该算法进行验证，结果表明经该算法降维后的数据集具有较好的聚类与分类性能。

关键词：振动与波；数据集降维；故障诊断；线性分块；LTSA；LLTSA

中图分类号：TP206 **文献标识码**：A DOI编码：10.3969/j.issn.1006-1335.2014.05.034

Dimension Reduction Method for Rotor Fault Dataset Based on LLTSA

YUAN De-qiang[1,2], *ZHAO Rong-zhen*[1,2]

(1. Key Laboratory of Digital Manufacturing Technology and Application, the Ministry of Education, Lanzhou University of Technology, Lanzhou 730050, China;
2. School of Mechanical & Electronic Engineering, Lanzhou University of Technology, Lanzhou 730050, China)

Abstract : Characteristic dataset dimension reduction is one of the key steps in mechanical fault intelligent diagnosis. However, it is difficult for the conventional dataset dimension reduction method to accurately extract the sensitive characteristic information, which can reflect the operation state of the rotor, from the high-dimensional nonlinear datasets. Thus, the precision of fault pattern recognition will be reduced. Although the local tangential space alignment (LTSA) algorithm can yield good dimension reduction effect for a part of the high-dimensional nonlinear datasets, it is not suitable for processing the high dimensional data sources with high-curvature distribution, sparse uneven distribution etc. In this study, combining the LTSA algorithm with the linear blocking method, the linear local tangential space alignment (LLTSA) algorithm is proposed. In the algorithm, the global and the local structures of the dataset are fully considered, and the data sample space is divided into a set of linear blocks so that the same data after the dimension reduction has a better clustering performance. This algorithm is verified by the time-domain characteristic dataset of the high-dimensional nonlinear rotor vibration data. The results show that the dataset after the dimension reduction algorithm has a good clustering and classification performance.

Key word : vibration and wave ; data set dimension reduction ; fault diagnosis ; linear block ; LTSA ; LLTS

收稿日期：2013-12-10

基金项目：国家自然科学基金项目(50875118，51165019)

作者简介：袁德强(1986-)，四川成都人，硕士，研究方向为旋转机械状态监测与故障诊断。
E-mail: 475072966@qq.com

通讯作者：赵荣珍(1960-)，山东枣庄人，教授，博士，博士生导师。
zhaorongzhen@lut.cn

随着机械工业的快速发展，计算机与网络的广泛使用，人们可以获得越来越多的机械信息量。伴随着表示信息的数据集维数越来越高，非平稳、非线性等时变特性越来越突出[1]。为得到高维数据集内在规律，需对其进行降维处理。目前比较成熟的数据降维方式以全局线性降维方法为主，如：主成分分析(PCA)、线性判别分析（LDA）、多维尺度分析

（MDS）[2, 3]等，以上方法均不适于非线性数据降维。针对此不足，文献 [4，5]运用了核主成分分析（KPCA），相对于上述方法，其能有效的对非线性数据集进行降维。但它得到数据低阶统计特性的同时忽略了其高阶统计特性中包含的重要特征信息。因此，寻找一种能高效准确的对高维非线性数据集进行降维的方法，即特征提取 [6]，是当前机械故障诊断亟需解决的问题。

数据降维尤其是非线性数据降维是数据挖掘的重要内容之一 [7, 8]，是近年来最实用的降维方法。以流形学习为代表，它旨在发现嵌入在高维非线性数据空间的低维光滑流形的内在结构，以便于对数据的深层理解和正确处理。目前其经典算法有Isomap、LLE、LTSA 等三种方法 [9]。这些算法都是通过找到每个数据点周围的局部信息及其性质，将流形几何结构通过非线性映射到一个低维的特征空间，同时保持原始数据的局部结构。LTSA 作为一种较好的非线性数据降维算法，能够有效地发现高维数据集的低维流形结构的整体嵌入坐标而得到广泛应用 [10-11]。

然而 LTSA 算法对邻近点数选择的依赖性较强，且不适于处理不均匀、高曲率数据。本文采用线性局部切空间排列算法(LLTSA)，较好的克服了LTSA 算法存在的缺点。它能有效的得到维数低、分类正确率高的主要敏感特征，达到高效简化故障原始特征集和区分故障类别的效果。

1　原理介绍

1.1　局部切空间排列算法(LTSA)

数据降维是将高维数据通过线性或非线性映射投影到低维空间来获得其低维表示 [12]。LTSA 作为非线性数据降维主要方法之一 [13]，其核心思想是利用样本中某个点及其邻域的切空间来表示这些数据点的局部几何性质，通过数据点在局部切空间上的投影，来获得局部低维坐标，然后将这些局部低维坐标经排列变换来构造全局低维坐标 [14]。该算法输入高维非线性数据集，目标输出低维数据。步骤如下：

Step1：选取邻域。对于样本中的某点 x_i，选取包含其自身在内的 k 个最邻近点作为邻域。

Step2：局部坐标线性拟合。求取点 x_i 处 d 维切空间的正交基 Q_i 以及邻域中每个数据点 x_{ij} 在切空间上的正交投影 $\theta_j^i = Q_i^{\mathrm{T}}(x_{ij} - \bar{x})$，其中 $\bar{x}$ 为 k 个邻域点的均值。

Step3：局部坐标全局整合。根据 n 个局部正交投影 $\Theta_i = [\theta_1^i, \theta_2^i, \cdots, \theta_n^i]$。通过最小化全局重构误差计算得到全局坐标 $T = [\tau_1, \tau_2, \cdots \tau_n]$。最小化全局重构误差算法

$$\sum_i \|E_i\|_2^2 = \sum_i \left\| T_i\left(I - \frac{1}{k_i} ee^{\mathrm{T}}\right) - L_i\Theta_i \right\|_2^2 \qquad (1)$$

LTSA 是一种假设全局非线性局部线性的数据降维方法。其能很好的对部分非线性数据进行降维，但其不能准确的把高曲率、不均匀分布的数据集划分成局部平滑线性数据块，难点就在于 k 值的选取。k 值过大，数据块线性程度降低；反之，计算复杂度增加。经研究分析，线性分块算法能将高维高曲率数据精确的分成线性块。合适的 k 值是建立在精确的线性分块基础上的。

1.2　线性分块算法

转子振动故障数据集不仅具有非线性特性，而且其数据分布的曲率不尽相同。曲率越高的区域，数据分布的非线性程度越高，则局部线性空间的分布越稠密；反之，数据线性化程度越好，局部线性空间的分布越稀疏。为得到最大线性数据块，可采用自适应划分方法，即线性分块算法。该算法的目的是构造最大近似线性数据块。其思路如下：

给定样本 $X = \{x_1, x_2, \ldots, x_N\}$，$x_i \in R^D$ 其中 x_i 为 D 维列向量，N 为样本数，数据点数为 $D \times N$。把这些数据点划分到不同的线性块中，且满足这些块是最大线性块。用 F_i 来表示最大线性块 i 的数据点数。其表示如下

$$X = \sum_{i=1}^{c} F_i \ ; \ F_i \bigcap F_j = \varphi, \ (i \neq j; \ i, \ j = 1, 2, \cdots, c) \qquad (2)$$

其中 $F_i = \{x_{1i}, x_{2i}, \cdots x_{N1}\}$，（$x_{1i}$ 是 x_1 中属于 F_i 的数据点数，后面依次类推），c 为线性块数。

算法如下：

Step1：初始化数据，设 $i = 1$，$F_{i-1} = \varphi$，$X_Y = \varphi$，$X_W = X$，其中 X_Y 表示已划分数据点，X_W 表示未划分数据点。

Step2：若 $X_W \neq \varphi$，则从中随机取一个数据点作为参照点，把其他所有未划分点与参照点的欧氏距离与测地距离之差 ρ 作为选取标准。若 $\rho \leqslant \theta$，将其与参照点划分在同一线性块中，记为 F_i。反之则弃之。其中 θ 为两点线性程度指标。

Step3：重复执行步骤 2，直到 $X_W = \varphi$ 为止，每执行一次步骤 2，i 值加 1。

该算法在保证数据块较高线性化程度的同时减少线性块数量，使计算量降低。从而更有利于处理高曲率、不均匀及大数据样本。若将其运用到非线

性数据降维中，降维效果更佳。

1.3 线性局部切空间排列算法（LLTSA）

LLTSA作为LTSA的局部最大线性扩展，充分考虑了高维样本数据集的局部信息和拓扑结构。其思路是将LTSA与自适应线性分块相结合，利用欧式距离与测地距离之差与预先给定的一个阀值θ比较来构造局部数据块，并通过几何摄动来严格控制非线性数据的局部线性特性[15]，使降维后特征信息的丢失小于原始的LTSA降维结果。LLTSA算法是将高维数据集通过线性分块算法划分成n个最大线性块。然后通过寻找每一个最大线性块的转换矩阵A，使样本数据$X=\{x_1, x_2,\ldots, x_m\}$，$x_i\in R^D$投影以后的样本变为$Y=(y_1, y_2,..., y_m)$，$y_i\in R^d$，$d<D$，且满足线性映射$Y=A^{\mathrm{T}}XH_m$，这里$H_m=I-ee^{\mathrm{T}}/\mathrm{m}$，$e$是元素都是1的$m$维的列向量。

设每个样本点x_i的最近k个点为$x_i=(x_{i1}, x_{i2},\cdots, x_{ik})$，$s_i$是只含0,1元素的选择矩阵，并满足$Y_i=Y_{si}$。为了保持局部线性最大化，达到自适应最佳线性程度，应使数据的重建误差E_i值最小。则可得到LLTSA的目标函数为

$$\arg_Y \min\sum_i\|E_i\|^2=\arg_Y\min\|YSW\|^2=$$

$$\arg Y\ \min \mathrm{tr}(\mathrm{YSWW^{\mathrm{T}}S^{\mathrm{T}}Y^{\mathrm{T}}}) \qquad (3)$$

其中$E_i=Y_iw_i=Y_{Si}w_i$，$S=[s_1,\ldots, s_m]$，并且$W=\mathrm{diag}(w_1,\ldots, w_m)$，$w_i$为$w_i=H_k(I-V_iv_i^{\mathrm{T}})$，其中$V_i$是与$X_iH_m$的第$d$个最大奇异值对应的右奇异向量。为了唯一确定$Y$，应限制$YY^{\mathrm{T}}=I_d$，其中$I_d$是单位矩阵。如果考虑映射$Y=A^{\mathrm{T}}XH_m$，那么目标函数就可以转化为如下

$$\arg Y\ \min \mathrm{tr}(A^{\mathrm{T}}XH_mBH_mA^{\mathrm{T}}A)$$

$$A^{\mathrm{T}}XH_mX^{\mathrm{T}}A=I_d \qquad (4)$$

其中$B=SWW^{\mathrm{T}}S^{\mathrm{T}}$。若使用拉格朗日乘子法，则最优的投影向量可通过求解如下广义特征值问题得到

$$XH_mBH_mX^{\mathrm{T}}a=\lambda XH_mX^{\mathrm{T}}a \qquad (5)$$

如果与特征值$\lambda_1\leqslant\lambda_2\leqslant\lambda_3\cdots\leqslant\lambda_d$对应的特征向量为$a_1, a_2,\ldots, a_d$，则LLTSA的转换矩阵为

$$A=[a_1, a_2,\ldots, a_d]$$

从而获得低维表示。

对降维结果进行聚类处理，可方便的比较各降维方法的降维效果。算法如下：

step 1 对于同类数据样本$X=\{x_1, x_2,\cdots x_n\}$；计算其样本均值$\bar{x}=\sum_{i=1}^{n}x_i/n$

step 2：计算$k_i=x_i-\bar{x}$；其中i=1，2$\cdots n$

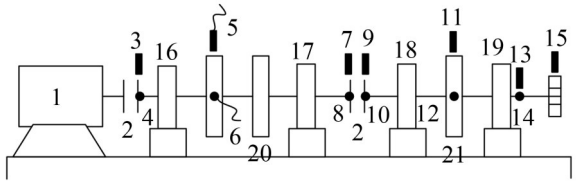
step 3：计算$R=\left|\sum_{i=1}^{n}k_i\right|$ （6）

降维结果通过聚类处理后可以充分说明同类数据的类内紧密程度，便于模式辨识。

2 实验与讨论

2.1 实 验

本文采用双跨转子实验台（如图1）采集的振动位移信号数据作为实验对象，在实验台13个通道上安装电涡流位移传感器，ch1—ch12是6对相互垂直安装用于测量6个截面水平（x）和竖直（y）方向上的振动位移信号。布置于实验台末端的ch 0用于测量转速。在本实验台上共模拟了两种常见故障（转子不平衡、转子不对中）和正常状况，并分别在不同转速与采样频率下对不同状态信号进行采集。最终选取采样频率为5 000 Hz，转速为3 000 r/min的振动信号进行分析。通过观察信号的时域波形发现第11通道的信号在不同状态下的区分度较为明显，因此选用第11通道的信号数据进行分析，截取长度设为5 120。由于时频域特征容易获得并蕴含丰富的状态特征[16]。因此本实验选取14个时域特征(表1)来进行分析。首先将3类状态下第11通道所采集得到的5 120个数据点划分成39个样本。够成一个3×256×39的原始高维矩阵（表2），然后对39个样本进行时域分析。得到3×39×14的特征矩阵（表3）。



1. 电动机、 2. 联轴器、 3. —15. 电涡流传感器
16.—20. 轴承、 21. —22. 质量盘

图1 实验装置简图

双跨转子系统获取的故障数据时域特征集呈现出较强的非线性特性。因此，要对转子系统做出正确的模式识别，需采用非线性降维方法对其进行处理。LLTSA方法不仅具备处理非线性数据的能力，而且还能很好的处理曲率不同、稀疏不均匀数据集。最后分别通过PCA，LTSA与LLTSA方法对特征矩阵进行降维，将得到的低维矩阵进行比较。如

表 1 时域特征参数

序号	时域特征参数名称	序号	时域特征参数名称
1	均值	8	峰峰值
2	均方根幅值	9	峰值指标
3	标准差	10	裕度指标
4	方差	11	波形指标
5	绝对均值	12	*C* 因素
6	偏度	13	均根方值
7	峭度	14	*L* 因素

图2。

2.2 讨 论

为直观显示降维结果，将LLTSA、PCA与LTSA输出的低维数据进行图像显示。经PCA降维得到的效果如图3：从图中可以看出对于正常状态时域特征，其降维效果较好，所有数据点都聚集在一起。而

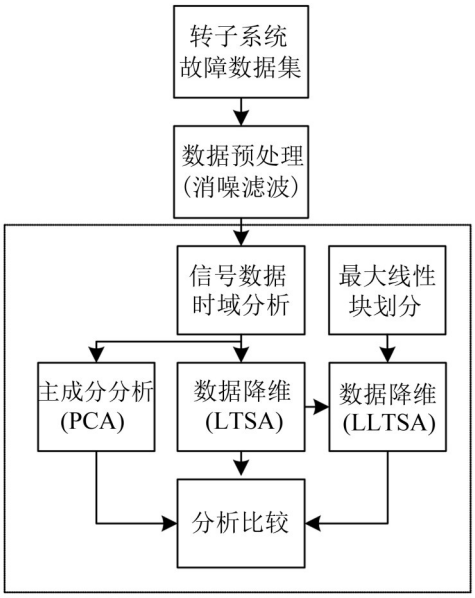


图 2 数据处理与比较

表 2 第11通道振动数据样本划分

数据类型	样本元素	样本1	样本2	样本3	样本4	……	样本36	样本37	样本38	样本39
	1	−0.391 9	0.016 7	0.388	−0.247 6	……	−0.194 1	−0.252 8	0.305 9	0.167
	2	−0.391 9	0.036 5	0.392 6	−0.273	……	−0.226 2	−0.233 9	0.323 6	0.135 3
正常	……	……	……	……	……	……	……	……	……	……
	255	0.378	−0.197 5	−0.258 3	0.271 3	……	0.269 9	0.221 2	−0.374	0.028 3
	256	0.382 9	−0.222 1	−0.241 6	0.288 8	……	0.288 1	0.196 2	−0.374	0.039
	1	−4.196 8	−3.847 7	−3.408 2	−4.445 8	……	−3.996 6	−3.425 3	−3.791 5	−4.172 4
	2	−4.213 9	−3.825 7	−3.430 2	−4.067 4	……	−3.969 7	−3.420 4	−3.833	−4.184 6
不平衡	……	……	……	……	……	……	……	……	……	……
	255	−3.388 7	−5	−4.108 9	−3.518 1	……	−3.740 2	−4.174 8	−3.662 1	−3.481 4
	256	−3.400 9	−4.011 2	−4.187	−3.508 3	……	−3.757 3	−4.309 1	−3.623	−3.542 5
	1	−3.293 5	−2.846 7	−2.790 5	−4.353	……	−2.824 7	−2.824 7	−3.291	−3.205 6
	2	−3.264 2	−2.941 9	−2.824 7	−3.476 6	……	−2.793	−2.827 1	−3.305 7	−3.173 8
不对中	……	……	……	……	……	……	……	……	……	……
	255	−2.734 4	−3.503 4	−3.046 9	−2.607 4	……	−3.256 8	−3.256 8	−2.600 1	−3.269
	256	−2.753 9	−3.491 2	−3.02	−2.585 4	……	−3.544 9	−3.234 9	−2.531 7	−3.120 1

表 3 第11通道振动数据时域特征

类型	样本编号	均值	均方根幅值	标准差	方差	……	波形指数	C 因素	均根方值	L 因素
正常	1	−0.000 1	0.269 2	0.269 7	0.072 8	……	1.118 5	3.026 3	0.219 3	3.715 2
	2	0.046 7	0.270 2	0.266 7	0.071 1	……	1.124 7	3.111 9	0.217 6	3.864 7
	……	……	……	……	……	……	……	……	……	……
	38	0.021 8	0.285 1	0.284 8	0.081 1	……	1.103	2.864 7	0.237 2	3.442 5
	39	−0.044 1	0.266 6	0.263 4	0.069 4	……	1.125	3.137 5	0.213	3.926 9
不平衡	1	−3.823 7	3.842	0.375 8	0.141 2	……	1.004 8	0.663 4	3.814 4	0.668 2
	2	−3.816 9	3.778 1	0.346 4	0.12	……	1.004 2	0.674 6	3.754 4	0.678 9
	……	……	……	……	……	……	……	……	……	……
	38	−3.860 7		0.350 3	0.122 7	……	1.004 1	0.728	3.852 6	0.732 6
	39	−3.816 9	3.876 5	0.360 7	0.130 1	……	1.004 4	0.608 8	3.808 7	0.612 8
不对中	1	−3.025 8	3.066 4	0.497 9	0.247 9	……	1.010 9	1.937 1	3.016 4	1.969 2
	2	−3.054 6	3.087 5	0.451	0.203 4	……	1.010 8	1.195 6	3.038 6	1.214 8
	……	……	……	……	……	……	……	……	……	……
	38	−3.084 2	3.109 4	0.396	0.156 8	……	1.008 2	0.922 6	3.072	0.933 8
	39	−3.020 6	3.048 1	0.409 5	0.167 7	……	1.009 1	1.072 5	3.007 8	1.086 9

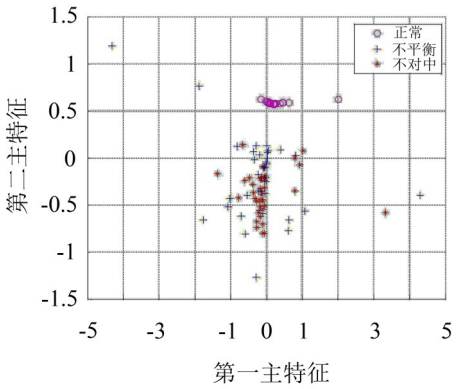


图 3 PCA降维效果图

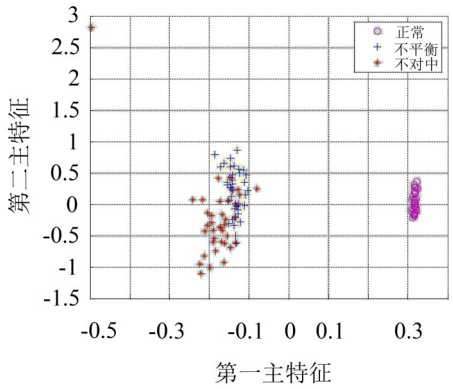


图 4 LTSA降维效果图

对于故障数据，其很难聚集到一处，两类故障数据相互嵌入。此降维方法能很好的处理线性数据；而对于转子故障非线性数据的处理效果却表现得极差。

经LTSA降维得到的效果如图4：从图中观察到正常数据能聚集在一起，而其它两类故障数据却没有完全分离开，k 值的选取不恰当是造成此结果的原因。时域特征的高曲率以及不均匀等特性没得到很好的处理。

经LLTSA方法得到的降维效果如图5：从图中可以看出该方法能将正常信号特征数据与故障特征数据以及故障特征数据之间很好的分开。几乎不存在重叠区。可见改进后的方法比传统的线性方法和原始的LTSA的效果都要好。

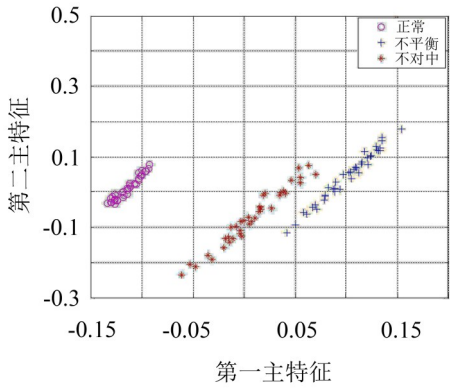


图 5 LLTSA降维效果图

为了进一步说明降维效果。本文采用聚类方法对降维后数据进行处理。所得结果如表4：

表4 三种数据聚类结果

方法	坐标	正常	不平衡	不对中
PCA	x	8.745 7	7.347 0	1.398 7
	y	21.999	8.165 8	13.833
LTSA	x	12.031	5.191 8	6.839 3
	y	9.699	7.743 5	8.713 4
LLTSA	x	1.212 7	3.896 5	0.335 0
	y	2.060 0	1.119 8	3.179 8

通过表4可以看出LLTSA降维后的数据具有更好的聚类效果。将更有利于转子系统的故障分类。

3 结 语

本文采用了一种改进的局部切空间排列算法（LLTSA），该算法在LTSA算法的基础上采用了几何摄动（自适应）获得了全局高维非线性数据集的局部最大线性块邻域，从而克服了LTSA方法中邻域k值的选择困难。采用LLTSA方法能直观的进行判别分类，通过转子信号时域特征降维后的敏感特征矢量验证了分类结果的正确性。在处理不均匀数据源和高曲率数据源时更稳定。转子振动数据集的降维分析验证了该方法较PCA与LTSA方法有更好的降维效果。因此，LLTSA适用于高曲率、不均匀数据集，提高了不平稳大数据的处理效率。

参考文献：

[1] 李文华. 改进的线性局部切空间排列算法[J]. **计算机应用**，2011，31(001)：247-249.

[2] Giri D, Rajendra Acharya U, Martis R J, et al. Automated diagnosis of coronary artery disease affected patients using LDA, PCA, ICA and discrete wavelet transform[J]. **Knowledge-Based Systems**, 2013, 37: 274-282.

[3] 梁胜杰，张志华，崔立林. 基于PCA方法的机械噪声数据处理[J]. **噪声与振动控制**，2010，3：032.

[4] 赵小强，王新明，王迎. 基于PCA与KPCA的TE过程故障检测应用研究[J]. **自动化仪表**，2011，32 (001)：8-12.

[5] 孙丽萍，陈果，陈立波，等. 基于KPCA的航空发动机滑油滤磨屑图像识别[J]. **机械科学与技术**，2010，29(6)：731-736.

[6] 夏均忠，苏涛，马宗坡，等. 基于EMD的滚动轴承故障特征提取方法[J]. **噪声与振动控制**，2013，2：027.

[7] 李锋，汤宝平，陈法法. 基于线性局部切空间排列维数化简的故障诊断[J]. **振动与冲击**，2012，31(13)：36-40.

[8] 杨剑，李伏欣，王珏. 一种改进的局部切空间排列算法[J]. **软件学报**，2005，16(9)：1584-1590.

[9] 杨庆，陈桂明，何庆飞，等. 局部切空间排列算法用于轴承早期故障诊断[J]. **振动、测试与诊断**，2012，5：028.

[10] 陈法法，汤宝平，苏祖强. 基于局部切空间排列与MSVM的齿轮箱故障诊断[J]. **振动与冲击**，2013，32 (5)：38-42.

[11] ZHANG Z, ZHA H. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment[J]. **Journal of Shanghai University (English Edition)**, 2004, 8 (4): 40 6-424.

[12] 万鹏，王红军，徐小力. 局部切空间排列和支持向量机的故障诊断模型[J]. **仪器仪表学报**，2012，33(12)：2789-2795.

[13] ZHANG T, YANG J, ZHAO D, et al. Linear local tangent space alignment and application to face recognition[J]. **Neurocomputing**, 2007, 70(7): 1547-1553.

[14] 顾艳春. 一种改进的局部切空间排列算法[J]. **计算机应用研究**，2013，30(3)：728-731.

[15] 杨安平，陈松乔，胡鹏. 基于几何距离摄动的局部切空间排列算法[J]. **计算机工程与应用**，2011，47(29)：168-170.

[16] 吴胜强，姜万录，刘思远. 基于小波包和KPCA的时频域故障检测方法[J]. **沈阳工业大学学报**，2011，33(2)：172-176.