基于局部质心均值最小距离鉴别投影的旋转机械故障数据降维分析研究

石明宽，赵荣珍

（兰州理工大学机电工程学院，甘肃 兰州 730050）

摘要：针对旋转机械故障特征集非线性强、维数过高导致分类困难的问题，提出一种基于局部质心均值最小距离鉴别投影（Local Centroid Mean Minimum-distance Discriminant Projection，LCMMDP）的故障数据集降维算法。该算法在考虑样本的内聚性和分离性的同时，能够保持样本局部几何结构信息，反映样本与局部质心均值之间的近邻关系。从多个角度提取机械振动信号的混合特征，构建原始高维特征集，通过LCMMDP提取出低维敏感特征子集，利用改进的基于局部均值与类均值的k-近质心近邻分类算法（k-nearest Centroid Neighbor Classification Based on Local Mean and Class Mean，KNCNCM）进行故障模式识别。所提方法集成了LCMMDP在维数约简和KNCNCM在模式识别的优势，可得到较高的故障识别准确率。分别使用一个双转子系统数据集和仿真数据集验证了该方法的有效性。

关键词：故障诊断；降维；局部质心均值；分类器；模式识别

中图分类号：TH 165$^+$3；TN911.7　　**文献标志码：**A　　**文章编号：**1004-4523(2021)02-0421-10

DOI：10.16385/j.cnki.issn.1004-4523.2021.02.023

引　言

旋转机械在现代机电系统中起着至关重要的作用，因此对旋转机械进行状态监测和故障诊断具有极其重要的意义[1]。为了尽可能多地获取故障信息，通常采用多个传感器进行多通道的监测，提取出每个通道的多域故障特征，必不可避免地导致了大量的冗余信息和高度相关的信息形成的“维数灾难”问题[2]。因此，如何消除高维数据中的冗余信息，使通过积累获得的海量故障数据资源拥有开发利用的价值，已成为当今机械信息技术所面临的基本问题。

伴随着大数据技术的快速发展，数据降维已成为数据科学研究领域关注的热点问题。典型的降维算法包括主成分分析（Principal Component Analysis，PCA）[3]与线性鉴别分析（Linear Discriminant Analysis，LDA）[4]等。其中，PCA试图寻找一个最大协方差线性投影矩阵，LDA通过最大化类间散度的同时最小化类内散度，寻找一个最佳投影矩阵；而PCA和LDA都是基于整体样本结构的降维算法，无法表征样本的局部流形结构。针对此问题，相关研究提出了流形学习算法，如局部保持投影（Locality Preserving Projection，LPP）[5]、局部线性嵌入（Locally Linear Embedding，LLE）[6]等算法。LPP是对传统拉普拉斯特征映射算法（Laplacian Eigenmap，LE）[7]进行线性化近似的结果，通过线性化之后的LPP能够以较小的计算损耗获取比较好的数据聚类效果。但LPP仅考虑了样本的局部结构，忽略了有利于分类的类别信息，无法更多地挖掘出高维数据的几何结构信息。针对这种不利的局面，将LDA与LPP算法的优势进行集成，提出了诸多改进算法，如局部Fisher判别分析（Local Fisher Discriminant Analysis，LFDA）[8]、边缘Fisher分析（Marginal Fisher Analysis，MFA）[9]、鉴别局部保持投影（Discriminant Locality Preserving Projection，DLPP）[10]、最小鉴别投影（Minimum-distance Discriminant Projection，MDP）[11]等算法。MDP通过引入类间相似度与类内相似度，不仅描述了样本与类中心的距离关系，同时反映出类间距与类内距的大小关系。但MDP和LDA在高维小样本问题中采用类均值会导致数据特征信息的丢失，不能更好地反映样本类信息[12]；另外MDP只考虑了样本与类中心之间的距离关系，忽视了样本点之间的局部近邻关系，无法有效地表征样本集的局部几何信息。针对MDP算法的局限性，本研究提出一种基于局部质心均值最小距离鉴别投影LCMMDP算法。LCMMDP利用样

收稿日期：2019-07-11；**修订日期：**2020-03-19

基金项目：国家自然科学基金资助项目（51675253）；兰州理工大学红柳一流学科建设项目

本与其近质心近邻点的均值间的距离设计了与MDP不同的相似性度量机制，欲从流形局部学习的角度重新定义局部类间相似度和局部类内相似度，充分利用了所有样本的局部几何信息和类别信息；另外，LCMMDP在计算过程中利用近质心近邻点的均值，能够有效地抑制噪声数据对流形学习的干扰，提高算法的鲁棒性。

为实现利用LCMMDP降维后得到的低维敏感特征矢量与故障类型间的准确识别，需选择一种精度高、稳定性好的分类器。基于局部均值的k-近质心近邻（Local Mean-based k-nearest Centroid Neighbor，LMKNCN）分类算法是Gou等[13]为克服离群点对分类性能的负面影响而提出的非参数分类器，其基本思想是用待测样本点到每一类的局部质心均值点的距离大小来指定待测样本的类别。LMKNCN只利用了未分类样本在每类里的近邻的局部均值信息，与类可分离性密切相关的类均值特性并未利用。针对上述问题，根据文献[14]的思想，本研究提出一种基于局部均值与类均值的k-近质心近邻分类方法KNCNCM，在类均值互不相同的情况下，既利用未分类样本在每类里的近质心近邻的局部均值信息，又利用类均值的整体信息进行分类的想法，可以提高LMKNCN的分类性能。

基于上述分析，本研究对LCMMDP与KNCNCM相结合的转子故障数据集降维和分类方法进行探讨，欲从海量数据中挖掘更充分的数据结构信息，为智能故障模式识别技术的发展提供了一种理论参考依据。

1 MDP算法的描述

MDP的基本思想为[11]：通过调节类内相似度与类间相似度权重参数的大小，使同类样本点对聚合，异类样本点对分离。

设一个高维数据集有n个D维向量$\boldsymbol{X}=\{\boldsymbol{x}_i|i=1,2,\cdots,n;\ \boldsymbol{x}_i\in\mathrm{R}^D\}$，分为$C$个类别。MDP定义的样本类内散度矩阵$\boldsymbol{S}_w$、类间散度矩阵$\boldsymbol{S}_b$分别如下所示：

$$\boldsymbol{S}_w=\sum_{c=1}^{C}\sum_{i=1}^{n_c}w_{ic}\left(\boldsymbol{x}_i^c-\boldsymbol{m}_c\right)\left(\boldsymbol{x}_i^c-\boldsymbol{m}_c\right)^{\mathrm{T}} \tag{1}$$

$$\boldsymbol{S}_b=\sum_{c=1}^{C}\sum_{i=1}^{n_c}\sum_{k=1,k\neq c}^{C}b_{ik}\left(\boldsymbol{x}_i^c-\boldsymbol{m}_k\right)\left(\boldsymbol{x}_i^c-\boldsymbol{m}_k\right)^{\mathrm{T}} \tag{2}$$

式中 w_{ic}，b_{ik}分别表示$\boldsymbol{x}_i^c$与类内中心点$\boldsymbol{m}_c$、第k（$k\neq c$）类中心点$\boldsymbol{m}_k$的相似度权重；$\boldsymbol{x}_i^c$为第c类的第i个样本；$\boldsymbol{m}_c=\dfrac{1}{n_c}\sum_{i=1}^{n_c}\boldsymbol{x}_i^c$，$\boldsymbol{m}_k=\dfrac{1}{n_k}\sum_{i=1}^{n_k}\boldsymbol{x}_i^k$分别为第$c$类、第$k$类的样本均值，相似度权值分别定义为：

$$w_{ic}=\exp\left(-\frac{\left\|\boldsymbol{x}_i^c-\boldsymbol{m}_c\right\|^2}{t}\right) \tag{3}$$

$$b_{ik}=\begin{cases}\dfrac{\exp\left(-\dfrac{\left\|\boldsymbol{x}_i^c-\boldsymbol{m}_k\right\|^2}{t}\right)}{\exp\left(-\dfrac{\left\|\boldsymbol{x}_i^c-\boldsymbol{m}_c\right\|^2}{t}\right)},\dfrac{\exp\left(-\dfrac{\left\|\boldsymbol{x}_i^c-\boldsymbol{m}_k\right\|^2}{t}\right)}{\exp\left(-\dfrac{\left\|\boldsymbol{x}_i^c-\boldsymbol{m}_c\right\|^2}{t}\right)}>\sigma & (4)\\ 0, \quad \text{others}\end{cases}$$

式中 $t>0$和$0\leqslant\sigma<1$是可调节参数。

则MDP的目标函数为

$$J_{\mathrm{MDP}}=\arg\max\frac{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{S}_b\boldsymbol{a}}{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{S}_w\boldsymbol{a}} \tag{5}$$

可以通过计算$\boldsymbol{S}_w^{-1}\boldsymbol{S}_b$的前$d$个最大特征值所对应的特征向量得到最优投影矩阵$\boldsymbol{A}=[\boldsymbol{a}_1,\ \boldsymbol{a}_2,\ \cdots,\ \boldsymbol{a}_d]$。

MDP是一种基于样本局部结构的算法，但MDP有以下几个缺点：①只反映了样本与类中心之间的距离关系，而忽视了样本间的近邻关系；②MDP利用类均值，在高维小样本问题中会导致故障特征信息丢失。针对上述问题，LCMMDP通过引入样本与其近质心近邻点的局部均值的相似度权重来刻画样本与局部质心均值的近邻关系，挖掘出更有利于分类的故障特征信息。

2 设计的LCMMDP降维算法

2.1 k-近质心近邻邻域的构造方法

传统的k-近邻方式构建邻域是基于样本点最近欧式距离的k个点作为近邻点，这种邻域构建方式的缺点是只考虑了样本点间的距离关系，无法表达每个样本点的局部结构特征。针对传统邻域构造方法的不足，本小节提出k-近质心近邻邻域构造方法。

利用k-近质心近邻构造样本$\boldsymbol{x}$的邻域时，根据距离准则和对称准则选取k个近质心近邻点，所选取的近质心近邻点不仅离样本$\boldsymbol{x}$较近，而且尽可能分布在样本$\boldsymbol{x}$的周围[13]。

因此，对于D维样本集$\boldsymbol{X}=\{\boldsymbol{x}_i|i=1,2,\cdots,n;\ \boldsymbol{x}_i\in\mathrm{R}^D\}$，样本$\boldsymbol{x}_i$在不同类别中的近质心近邻点可通过以下两步迭代获得：

（1）寻找$\boldsymbol{x}_i$的第一个近质心近邻点，即离$\boldsymbol{x}_i$最近的点，记$\boldsymbol{x}_1^{\mathrm{NCN}}$；

（2）寻找$\boldsymbol{x}_i$的第k个近质心近邻点$\boldsymbol{x}_k^{\mathrm{NCN}}$（$k\geqslant2$），计算$\boldsymbol{X}_i$（$\boldsymbol{X}=\{\boldsymbol{X}_i|i=1,\ 2,\ \cdots,\ C;\ \boldsymbol{X}_i\in\mathrm{R}^D\}$，$\boldsymbol{X}_i$为每一类样本集中所选取的子集）中的每个点与所得的前$k-1$个近质心近邻点的质心点，然后计算每个质心点到样本$\boldsymbol{x}_i$的距离，选择到样本$\boldsymbol{x}_i$距离最近的质心

点所对应 $\boldsymbol{X}_i$ 中的那个数据点作为第 k 个近质心近邻点 $\boldsymbol{x}_k^{\mathrm{NCN}}$。

图 1 描述了两种邻域构建方法的区别与联系，k-近邻方式得到的近邻点用圈（○）表示，k-近质心近邻方式得到的近邻点用星（☆）表示。由图中可以看到 k-近邻与 k-近质心近邻得到的第一个近邻点相同，这是因为第一个近邻点都是以最小距离选取的；而其他近邻点不一定相同，k-近质心近邻构建的邻域半径较大，而且其近邻点均匀分布在样本点周围。

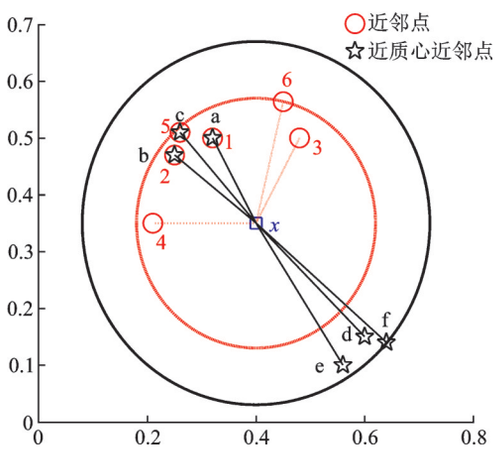


图 1　k-近质心近邻与 k-近邻的区别

Fig. 1　Difference between k-nearest centroid neighbor and k-nearest neighbor

2.2　确定的目标函数

针对 MDP 的不足，LCMMDP 算法在定义局部类内散度、局部类间散度时，采用局部学习的方式，使得 LCMMDP 充分利用类别信息和局部结构信息。

对于样本 $\boldsymbol{x}_i$，选择与其类别相同并且距离最近的 k_1 个近质心近邻点组成 $\boldsymbol{x}_i$ 的同类局部邻域，记为 $\boldsymbol{X}_{ij}^c=\left\{\boldsymbol{x}_{ij}^{\mathrm{NCN}}\middle|j=1,2,\cdots,k_1;\ \boldsymbol{x}_{ij}^{\mathrm{NCN}}\in\mathbf{R}^D\right\}$；然后在此基础上定义一个基于同类局部邻域的局部类内散度 $\widetilde{S}_w'$ 如下

$$\begin{aligned}\widetilde{S}_w'&=\sum_{c=1}^{C}\sum_{i=1}^{n_c}w_{ic}(\boldsymbol{y}_i^c-\widetilde{\boldsymbol{m}}_c^{\mathrm{NCN}})^2=\\&\sum_{c=1}^{C}\sum_{i=1}^{n_c}w_{ic}(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_i^c-\boldsymbol{a}^{\mathrm{T}}\boldsymbol{m}_c^{\mathrm{NCN}})^2=\\&\boldsymbol{a}^{\mathrm{T}}\left[\sum_{c=1}^{C}\sum_{i=1}^{n_c}w_{ic}(\boldsymbol{x}_i^c-\boldsymbol{m}_c^{\mathrm{NCN}})\cdot\right.\\&\left.(\boldsymbol{x}_i^c-\boldsymbol{m}_c^{\mathrm{NCN}})^{\mathrm{T}}\right]\boldsymbol{a}=\boldsymbol{a}^{\mathrm{T}}\boldsymbol{S}_w'\boldsymbol{a}\end{aligned}\tag{6}$$

式中　$\boldsymbol{S}_w'$ 为类内散度矩阵；$\boldsymbol{m}_c^{\mathrm{NCN}}=\dfrac{1}{k_1}\sum_{j=1}^{k_1}\boldsymbol{x}_{ij}^{\mathrm{NCN}}$ 为样本 $\boldsymbol{x}_i$ 的 k_1 个同类近质心近邻点的均值，$\widetilde{\boldsymbol{m}}_c^{\mathrm{NCN}}$ 为 $\boldsymbol{m}_c^{\mathrm{NCN}}$ 投影后的数据。

对第 c 类样本 $\boldsymbol{x}_i^c$，在第 k 类样本中（$c\neq k$）选择与 $\boldsymbol{x}_i^c$ 类别不同且距离最近的 k_2 个近质心近邻点，组成在第 k 类中样本 $\boldsymbol{x}_i^c$ 的异类局部邻域（共有 $C-1$ 个异类局部邻域），记为 $\boldsymbol{X}_{ij}^k=\left\{\boldsymbol{x}_{ij}^{\mathrm{NCN}}\middle|j=1,2\cdots,k_2;\ \boldsymbol{x}_{ij}^{\mathrm{NCN}}\in\mathbf{R}^D\right\}$；然后在此基础上定义一个基于异类局部邻域的局部类间散度 $\widetilde{S}_b'$ 如下

$$\begin{aligned}\widetilde{S}_b'&=\sum_{c=1}^{C}\sum_{i=1}^{n_c}\sum_{k=1,k\neq c}^{C}b_{ik}\left(\boldsymbol{y}_i^c-\widetilde{\boldsymbol{m}}_k^{\mathrm{NCN}}\right)^2=\\&\sum_{c=1}^{C}\sum_{i=1}^{n_c}\sum_{k=1,k\neq c}^{C}b_{ik}\left(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_i^c-\boldsymbol{a}^{\mathrm{T}}\boldsymbol{m}_k^{\mathrm{NCN}}\right)^2=\\&\boldsymbol{a}^{\mathrm{T}}\left[\sum_{c=1}^{C}\sum_{i=1}^{n_c}\sum_{k=1,k\neq c}^{C}b_{ik}(\boldsymbol{x}_i^c-\boldsymbol{m}_k^{\mathrm{NCN}})\cdot\right.\\&\left.(\boldsymbol{x}_i^c-\boldsymbol{m}_k^{\mathrm{NCN}})^{\mathrm{T}}\right]\boldsymbol{a}=\boldsymbol{a}^{\mathrm{T}}\boldsymbol{S}_b'\boldsymbol{a}\end{aligned}\tag{7}$$

式中　$\boldsymbol{S}_b'$ 为类间散度矩阵；$\boldsymbol{m}_k^{\mathrm{NCN}}=\dfrac{1}{k_2}\sum_{j=1}^{k_2}\boldsymbol{x}_{ij}^{\mathrm{NCN}}$ 为样本 $\boldsymbol{x}_i^c$ 在第 k 类样本中的 k_2 个近质心近邻点的均值，$\widetilde{\boldsymbol{m}}_k^{\mathrm{NCN}}$ 为 $\boldsymbol{m}_k^{\mathrm{NCN}}$ 投影后的数据。

类似于 MDP，定义 LCMMDP 准则函数为

$$J=\arg\max\frac{\widetilde{S}_b'}{\widetilde{S}_w'}=\arg\max\frac{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{S}_b'\boldsymbol{a}}{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{S}_w'\boldsymbol{a}}\tag{8}$$

为求满足 LCMMDP 准则函数的投影向量，只需求解 $(\boldsymbol{S}_w')^{-1}\boldsymbol{S}_b'$ 的 d 个最大特征值所对应的 d 个特征向量 $\boldsymbol{a}_1,\cdots,\boldsymbol{a}_d$，记投影矩阵 $\boldsymbol{A}=[\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_d]$。

2.3　权重的定义

w_{ic}，b_{ik} 分别表示 $\boldsymbol{x}_i^c$ 与类内局部质心均值点 $\boldsymbol{m}_c^{\mathrm{NCN}}$、第 k（$k\neq c$）类类间局部质心均值点 $\boldsymbol{m}_k^{\mathrm{NCN}}$ 的相似度权重，其权值分别定义为：

$$w_{ic}=\exp\left(-\frac{\left\|\boldsymbol{x}_i^c-\boldsymbol{m}_c^{\mathrm{NCN}}\right\|^2}{t}\right)\tag{9}$$

$$b_{ik}=\begin{cases}\dfrac{\exp\left(-\dfrac{\left\|\boldsymbol{x}_i^c-\boldsymbol{m}_k^{\mathrm{NCN}}\right\|^2}{t}\right)}{\exp\left(-\dfrac{\left\|\boldsymbol{x}_i^c-\boldsymbol{m}_c^{\mathrm{NCN}}\right\|^2}{t}\right)}, & \dfrac{\exp\left(-\dfrac{\left\|\boldsymbol{x}_i^c-\boldsymbol{m}_k^{\mathrm{NCN}}\right\|^2}{t}\right)}{\exp\left(-\dfrac{\left\|\boldsymbol{x}_i^c-\boldsymbol{m}_c^{\mathrm{NCN}}\right\|^2}{t}\right)}>\sigma\\0, & \text{others}\end{cases}\tag{10}$$

式中　$\left\|\boldsymbol{x}_i^c-\boldsymbol{m}_c^{\mathrm{NCN}}\right\|$ 为 $\boldsymbol{x}_i^c$ 与 $\boldsymbol{m}_c^{\mathrm{NCN}}$ 之间的欧式距离；本文中 t 取所有样本之间欧式距离均值的平方[15]。

2.4　LCMMDP 算法步骤规划

LCMMDP 算法具体步骤如下：

步骤 1：根据近质心近邻点数量和类别信息分别构建同类近邻图和异类近邻图。

步骤 2：根据式（9）计算样本 $\boldsymbol{x}_i^c$ 与类内局部质心均值点 $\boldsymbol{m}_c^{\mathrm{NCN}}$ 之间的相似度 w_{ic}；根据式（10）计算样

本 $\boldsymbol{x}_i^c$ 与第 $k(k\neq c)$ 类类间局部质心均值点之间的相似度 b_{ik}。

步骤 3：分别根据式(6)，(7)计算类内散度 $\widetilde{S}'_w$，类间散度 $\widetilde{S}'_b$。

步骤 4：通过式(8)计算 $\left(\boldsymbol{S}'_w\right)^{-1}\boldsymbol{S}'_b$ 前 d 个最大特征值及其对应的特征向量，构建投影矩阵 $\boldsymbol{A}=[\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_d]$，并将初始样本集 $\boldsymbol{X}$ 通过矩阵 $\boldsymbol{A}$ 进行降维投影，得到映射后的低维特征子集 $\boldsymbol{Y}=\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}$。

3 提出的KNCNCM分类算法

基于局部均值与类均值的k-近质心近邻分类算法(KNCNCM)的基本原理是计算待测样本 $\boldsymbol{x}$ 到每一类 k 个近质心近邻点的局部均值的距离以及测试样本 $\boldsymbol{x}$ 到对应类的均值的距离，用两个距离的组合距离来判断待测样本 $\boldsymbol{x}$ 的类别。

设 $T=\left\{\boldsymbol{x}_i\middle|i=1,2,\cdots,N;\ \boldsymbol{x}_i\in\mathrm{R}^d\right\}$ 是一组给定的训练样本集，该样本集由 d 个属性，N 个样本组成。它包含 C 个类别，其类别标签分别为 $c_1,c_2,\cdots,c_C$，$\boldsymbol{T}_i=\left\{\boldsymbol{x}_{ij}\middle|j=1,2,\cdots,N_i;\ \boldsymbol{x}_{ij}\in\mathrm{R}^d\right\}$ 是 T 中所有属于类 c_i 的样本集，N_i 为类 c_i 的样本个数，$\boldsymbol{x}$ 是一个待测样本，KNCNCM的分类步骤如下：

(1)从第 c_i 类的训练样本集 T_i 中为每一个待测样本 $\boldsymbol{x}$ 选取 k 个近质心近邻点，记为

$$T_{ik}^{\mathrm{NCN}}(\boldsymbol{x})=\left\{\boldsymbol{x}_{ij}^{\mathrm{NCN}}\middle|j=1,2,\cdots,k;\ \boldsymbol{x}_{ij}^{\mathrm{NCN}}\in\mathrm{R}^m\right\}$$

(2)计算从第 c_i 类中选取待测样本 $\boldsymbol{x}$ 的 k 个近质心近邻点的质心均值点，记为

$$\bar{\boldsymbol{u}}_{ik}^{\mathrm{NCN}}=\frac{1}{k}\sum_{j=1}^{k}\boldsymbol{x}_{ij}^{\mathrm{NCN}}\tag{11}$$

(3)计算 $\boldsymbol{x}$ 到第 c_i 类的局部质心均值 $\bar{\boldsymbol{u}}_{ik}^{\mathrm{NCN}}$ 的欧式距离 $d_{ik}(\boldsymbol{x},\bar{\boldsymbol{u}}_{ik}^{\mathrm{NCN}})$，即

$$d_{ik}\left(\boldsymbol{x},\bar{\boldsymbol{u}}_{ik}^{\mathrm{NCN}}\right)=\sqrt{\left(\boldsymbol{x}-\bar{\boldsymbol{u}}_{ik}^{\mathrm{NCN}}\right)\left(\boldsymbol{x}-\bar{\boldsymbol{u}}_{ik}^{\mathrm{NCN}}\right)^{\mathrm{T}}}\tag{12}$$

(4)计算第 c_i 类的均值 $\boldsymbol{u}_i$

$$\boldsymbol{u}_i=\frac{1}{N_i}\sum_{j=1}^{N_i}\boldsymbol{x}_{ij}\tag{13}$$

(5)计算 $\boldsymbol{x}$ 到 $\boldsymbol{u}_i$ 的欧式距离 $d_i(\boldsymbol{x},\boldsymbol{u}_i)$

$$d_i\left(\boldsymbol{x},\boldsymbol{u}_i\right)=\sqrt{\left(\boldsymbol{x}-\boldsymbol{u}_i\right)\left(\boldsymbol{x}-\boldsymbol{u}_i\right)^{\mathrm{T}}}\tag{14}$$

(6)用下式计算组合距离 d，即

$$d=d_{ik}(\boldsymbol{x},\bar{\boldsymbol{u}}_{ik}^{\mathrm{NCN}})+w\times d_i(\boldsymbol{x},\boldsymbol{u}_i)\tag{15}$$

式中 w 为距离加权系数，它反映了类均值对分类结果的影响程度。该值越大，说明对分类结果影响越大。w 的取值为 $0\leqslant w\leqslant 1$，本文根据下式取值[14]

$$w=1.25^{-(i-1)}\ \text{or}\ w=0,\ i=1,2,\cdots,41\tag{16}$$

(7)最小组合距离对应的均值点所属的类别即为待测样本点 $\boldsymbol{x}$ 的类别，即

$$m=\arg\min_{c_i}\ d\tag{17}$$

式中 m 即为待测样本 $\boldsymbol{x}$ 的分类结果。

4 基于LCMMDP与KNCNCM融合的故障诊断方法

为消除高维数据中的冗余信息，解决故障特征集维数过高的问题，本文提出了LCMMDP维数约简与KNCNCM分类器相结合的故障数据集分类方法。用LCMMDP算法对初始特征集进行维数约简得到低维敏感特征子集，然后利用KNCNCM分类算法进行故障模式识别。基于本文所提方法设计的故障流程如图2所示。

具体实现过程步骤如下：

输入：初始数据集 $\boldsymbol{X}=\left\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\right\}$，LCMMDP算法近邻值 k_1,k_2，低维空间维数 d，调节参数 σ。

输出：低维敏感特征子集 $\boldsymbol{Y}$，投影矩阵 $\boldsymbol{A}$。

步骤 1：提取振动信号的11个时域特征参数(p_1-p_{11})和10个频域特征参数(p_{12}-p_{21})组成初始特征集，如表1所示。

步骤 2：对初始特征集进行归一化处理后分为训练样本集 $\boldsymbol{X}_1$ 和测试样本集 $\boldsymbol{X}_2$ 两部分，将 $\boldsymbol{X}_1$ 输入LCMMDP算法中进行维数约简，通过对所构造的准则函数的求解，得到映射矩阵 $\boldsymbol{A}$。用 $\boldsymbol{A}$ 对 $\boldsymbol{X}_1,\boldsymbol{X}_2$ 进行特征投影得到低维敏感特征集 $\boldsymbol{Y}_1,\boldsymbol{Y}_2$。

步骤 3：将低维敏感特征集 $\boldsymbol{Y}_1$，$\boldsymbol{Y}_2$ 输入到KNCNCM分类器中，得到测试样本的故障类别。

5 实验验证及结果分析

为验证所提方法的有效性，选取UCI数据库中的Iris仿真数据集[16]与转子故障数据集进行验证。

5.1 Iris仿真数据验证实验

Iris数据集以鸢尾花的特征作为数据来源，数据集包含150个样本，分为3类，每类50个数据，每个数据包含4个属性。本文设置每类数据前20个为训练样本，后30个为测试样本。为验证LCMMDP的可行性，将仿真数据集经LDA，MDP，LCMMDP降维，目标维数为2=3-1。图3中：(a)，(b)，(c)，(d)分别为原始数据与LDA，MDP，LCMMDP降维后测试样本的可视化结果图。

由图3可知：原始数据集的第二类与第三类存在严重的混叠；经LDA，MDP，LCMMDP降维后的

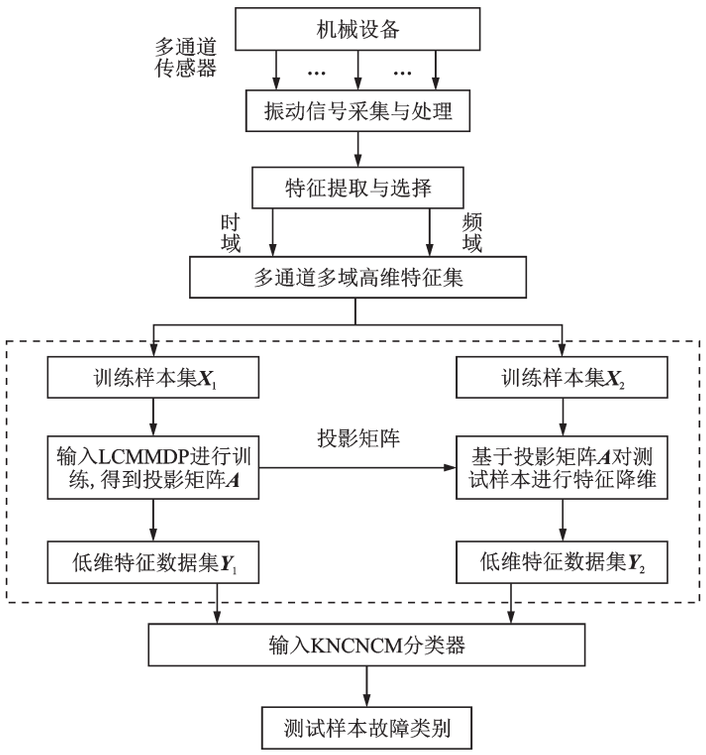


图 2　故障诊断流程图

Fig. 2　Procedure of fault diagnosis

表 1　特征参数

Tab. 1　Characteristic parameters

序号	特征名称	序号	特征名称
p_1	均值	p_{13}	频率方差
p_2	均方根值	p_{14}	频域峭度
p_3	方根幅值	p_{15}	频谱一阶重心
p_4	绝对平均值	p_{16}	频谱二阶重心
p_5	标准差	p_{17}	频谱二阶距
p_6	峰峰值	p_{18}	反映主频带位置的变化特征
p_7	波形指标		
p_8	峰值指标	p_{19}	参数
p_9	脉冲指标	p_{20}	反映频谱的集中或分散程度
p_{10}	裕度指标		
p_{11}	峭度指标	p_{21}	特征参数
p_{12}	均值频率		

注：

$$p_{18}=\sqrt{\frac{\sum_{k=1}^{K}f_k^4 s(k)}{\sum_{k=1}^{K}f_k^2 s(k)}};\ p_{19}=\frac{\sum_{k=1}^{K}f_k^2 s(k)}{\sqrt{\sum_{k=1}^{K}s(k)\sum_{k=1}^{K}f_k^4 s(k)}};$$

$$p_{20}=\frac{p_{16}}{p_{15}};\ p_{21}=\frac{\sum_{k=1}^{K}\left(f_k-p_{15}\right)^3 s(k)}{Kp_{16}^3};\ s(k)\text{为信号}x(n)\text{的频谱，}k=1,\cdots,K;K\text{为频谱线数；}f_k\text{为第}k\text{条谱线的频率值}$$

各类数据基本上可以分开，但LDA，MDP的类内比较分散，而LCMMDP降维后的数据类内比较聚集，类间比较分散。将降维得到的低维特征子集输入KNCNCM分类器中进行模式识别，得到测试样本识别准确率如表2所示。结合图3与表2可知：LCMMDP可得到较好的降维效果和较高的识别准确率，相比于其他算法有明显的优势。

5.2　转子故障模拟实验

为进一步验证本研究所提方法的可行性、有效性，本文通过如图4所示的一套双转子系统实验台进行研究分析。在转子系统中，轴1长415 mm，轴2长350 mm，直径都为15 mm，轴1上布置两个质量盘，轴2上布置一个质量盘，转子被4个轴承支承并被分隔为双跨结构形式。该双转子系统的第一临界转速约为2500 r/min，失稳转速约为5000 r/min，电机最高转速可达12000 r/min。12个电涡流传感器布置在转子系统的6个关键面处相互垂直方位上，通过不同方位采集转子系统的振动信号，第13个传感器安置在电机端用来采集转速信号。

转子系统中常见故障的有：转子不对中、质量不平衡、转子裂纹、转子弯曲、轴承座松动、动静碰摩等。本研究在转速3000 r/min，采样频率为5000 Hz的情况下，模拟了4种典型故障（转子不对中、质量不平衡、动静碰磨、轴承座松动）转动实验及正常状态转动实验。通常，转子不对中主要是由各个转子的轴心存在偏差造成的；质量不平衡是由转子的几何中心与质量中心存在偏心造成的；动静碰摩主要

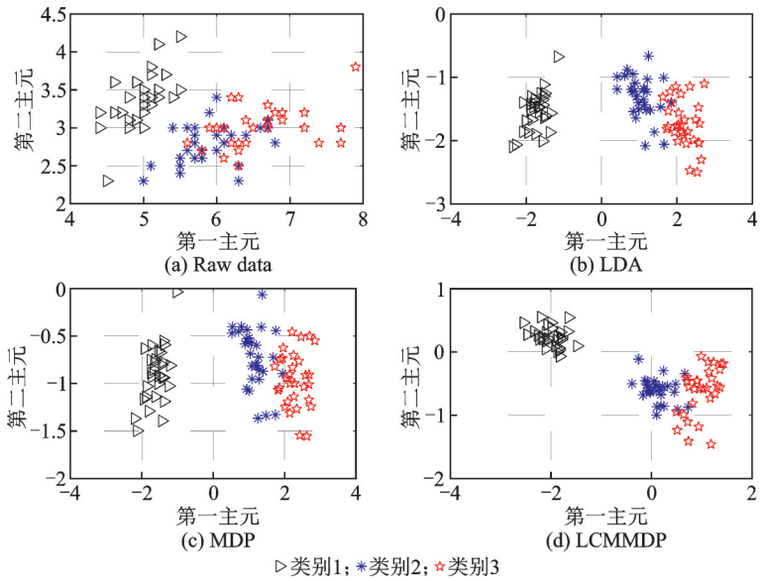


图 3 仿真数据特征分布图

Fig. 3 Feature maps of simulation data

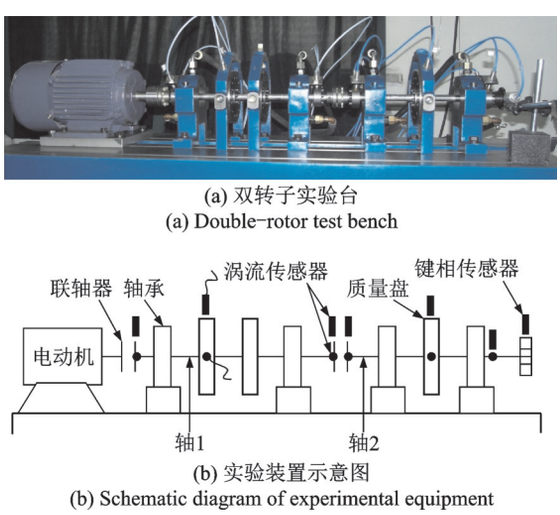
表 2 各降维方法的识别准确率

Tab. 2 Methods of dimension reduction of recognition accuracy

分类器	不同降维方法下的平均识别率/ %			
	原始	LDA	MDP	LCMMDP
KNCNCM	75.56	94.44	95.56	97.78

由转子不对中和转子不平衡造成转子与固定件接触引起的振动;轴承座松动是指转子系统接合面存在间隙或联结刚度不足造成机械阻尼偏低、振动过大。

采集每种状态的数据样本 80 组,其中 30 组作为训练样本,50 组作为测试样本。针对每个通道的传感器采集的振动信号分别提取时域、频域共 21 个特征参数,12 个通道总共得到 12×21=252 个特征参数。



(a) 双转子实验台
(a) Double-rotor test bench

(b) 实验装置示意图
(b) Schematic diagram of experimental equipment

图 4 双转子实验台示意图

Fig. 4 Schematic diagram of double-span rotor test bench

5.2.1 参数设定

本文需要设定的参数包括:流形学习 LCMMDP 中的近邻值 k_1,k_2,低维空间的维数 d,调节参数 σ 和分类器 KNCNCM 中的近邻值 k 及距离加权系数 w。现今有关流形学习的参数选择仍然没有统一的标准,通常将低维空间的维数 d 设定为样本类别数减 1,在本文中 $d=4$;近邻值的大小通常满足大于低维空间的维数 d、小于个样本的样本数 N_i ($i=1,2,\cdots,C$),即 $d<k<N_i$,在本文应该满足 $4<k<30$[17]。在实验中,k_1,k_2 的搜索范围为 $\{5,6,\cdots,29\}$,σ 的搜索范围为 $\{0,0.1,\cdots,0.9\}$,通过重复实验在设定的参数范围内循环搜索选取最优参数,当 $k_1=10$,$k_2=9$,$\sigma=0.7$ 时,LCMMDP 具有较好的特征集可分性。用交叉验证方法获得 KNCNCM 的近邻值 k 和 w,当 $k=7$,$w=1.25^{-(6-1)}=0.32768$ 时,故障识别准确率达到最大。

5.2.2 特征数据集的可分性分析

为了验证本文所提 LCMMDP 算法的可行性,选择与 LPP,LDA,MMC,MNMP,MDP 等降维算法进行比较。并将 6 种算法记为 A1,A2,A3,A4,A5,A6。前三个主元的低维嵌入结果如图 5 所示(图中“◇”、“*”、“☆”、“▷”、“○”分别代表转子不对中、质量不平衡、动静碰摩、轴承座松动和正常状态)。

从图 5 可以看出,LPP 的聚类效果最差,其中质量不平衡、动静碰摩、轴承座松动和正常状态四类特征之间离的较近,不易区分;MMC,MNMP 的同类特征太过分散,类内距离过大;LCMMDP 的聚类效果最好,不同类型特征之间完全分离,各类数据清晰

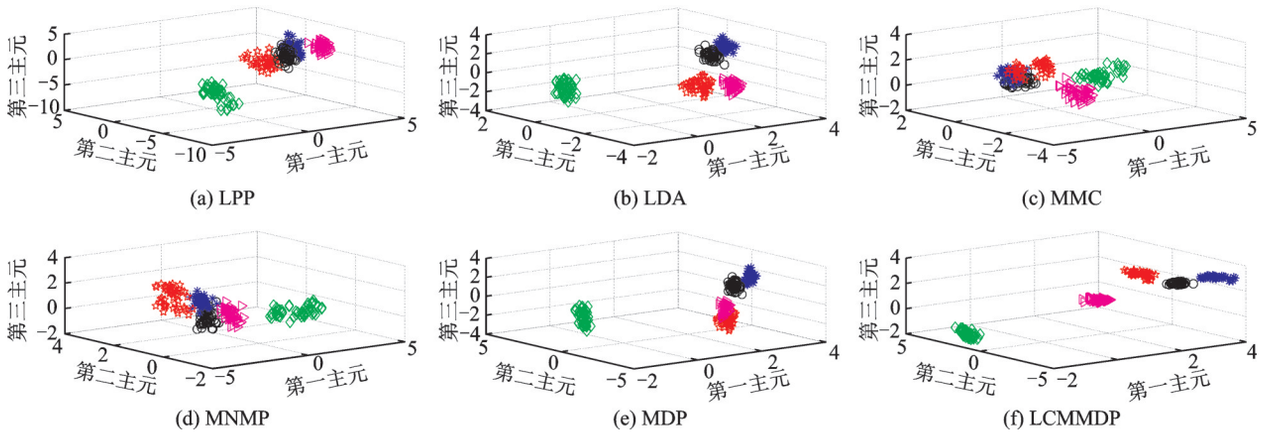


图 5　测试样本基于不同方法的降维结果

Fig. 5　Test sample based on different dimension reduction method results

可见，相同类型聚集成团；LPP，LDA，MMC，MNMP，MDP 五种算法的不同类型特征之间都存在着一定的混叠，无法有效地区分。

为了进一步说明本文所提 LCMMDP 算法的可行性，考察降维后的特征集的可分性，引入类间类内距离的可分性参数[18]对测试样本集的分类与聚类程度给予量化评价。设降维后的低维特征子集的 C 类样本数据为 $x_{ri}(r=1,2,\cdots,C;i=1,2,\cdots,N)$，其聚类中心为 $\bar{x}_r=\frac{1}{N}\sum_{i=1}^{N}x_{ri}\quad(r=1,2,\cdots,C)$。低维空间特征集的类间距离、类内距离和可分性参数分别定义为：

$$S_b=\frac{2}{C(C-1)}\sum_{r=1}^{C-1}\sum_{s=r+1}^{C}\left\|\bar{x}_r-\bar{x}_s\right\| \tag{18}$$

$$S_w=\frac{1}{C}\sum_{r=1}^{C}\frac{1}{N}\sum_{i=1}^{N}\left\|x_{ri}-\bar{x}_r\right\| \tag{19}$$

$$\rho=\frac{S_b}{S_w} \tag{20}$$

由定义可知，类间距 S_b 反映了各类别之间的分离度，类内距 S_w 体现了每类样本分布的紧凑程度。低维空间特征集的类间距离越大，类内距离越小，则可分性参数值 ρ 就越大，特征集的可分性就越好。

按式(20)计算各降维算法降维后的特征集可分性参数值，如表 3 所示。从表 3 可以看出，MMC 特征集的可分性参数最小，只有 12.4167；LCMMDP 特征集的可分性参数最大，可高达 100.3742。综合图 5 和表 3 可知，相比于其他 5 种降维算法，LCMMDP 聚类效果最好，可分性参数最大，说明 LCMMDP 具有较好的降维效果，在维数约简中有明显的优势。

表 3　各类方法降维后的类间类内可分性参数

Tab. 3　The separability parameters between classes and within classes after dimensionality reduction of various methods

降维算法	可分性参数 ρ
LPP	26.1011
LDA	18.9978
MMC	12.4167
MNMP	17.2676
MDP	17.9064
LCMMDP	100.3742

5.2.3　故障分类效果

为了量化 6 种降维算法的降维效果，将各个算法降维后的低维敏感特征集输入 KNCNCM 分类器中进行故障模式识别，得到的识别率如表 4 所示。

表 4　各类降维方法的识别准确率

Tab. 4　Classification identification method recognition accuracy

降维算法	各个状态类别的诊断准确率/%					平均识别准确率/%
	不对中	不平衡	碰磨	松动	正常	
LPP	100	24	100	100	92	83.2
LDA	100	97	94	100	79	94.0
MMC	100	92	88	98	58	87.2
MNMP	100	100	90	100	72	92.4
MDP	100	100	96	100	82	95.6
LCMMDP	100	100	100	100	100	100.0

结合表3和表4可以看出：①LPP的识别准确率最低，这是因为LPP是无监督算法，只侧重于局部几何结构信息的提取，而没有考虑类判别信息，导致故障特征解耦不完全，故障特征间仍存在混叠。②LDA，MMC，MNMP相较于LPP的识别率较好，是因为这三种算法均利用了样本的类判别信息，同时考虑了样本的分离性和内聚性。然而，LDA，MMC主要侧重于分析类判别信息，忽视了样本集的局部几何结构信息，导致大量的有用故障特征信息丢失；MMC，MNMP因为不存在小样本问题，没有利用PCA预维数约简处理，因此无法有效的除去空间中的噪声和冗余信息，导致提取的特征信息无法有效识别故障类别，同时降低了特征集的可分性。③LCMMDP的识别准确率要远高于MDP及其他四种降维算法，是因为LCMMDP将样本的局部几何信息有效地融入到维数约简过程，实现了类判别信息与样本集局部几何结构信息的有效结合。另外，LMMDP利用了局部质心均值，有效地克制了原始特征信息的丢失，一定程度上抑制噪声数据对算法的影响，在挖掘故障样本数据集中蕴含的故障信息的同时实现对故障的有效解耦，可得到最有辨识力的低维特征子集，提高故障特征集的可分性。

为了进一步验证LCMMDP算法的适用性，本文在选取不同的训练样本数量和测试样本数量的情况下，上述6种算法降维得到的低维敏感特征经KNCNCM分类器进行故障模式识别的结果如图6所示。

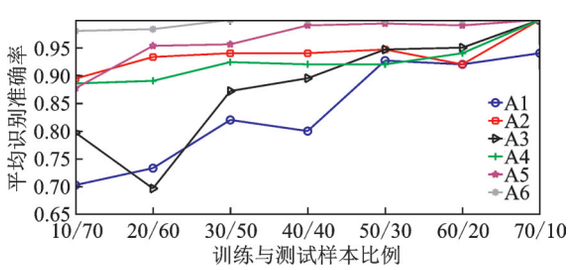


图6 不同训练样本数对应的平均识别准确率

Fig. 6 The average recognition accuracy of different training samples

图6表明，整体上各降维算法的识别准确率都随训练样本的增加而增加。由于LCMMDP充分利用了所有样本的局部几何信息和类别信息，故识别准确率一直较高，稳定性最好；LPP是无监督算法，在少量训练样本情况下，无法有效地保持局部结构信息，识别准确率较低；MMC是一种基于样本整体结构的降维算法，随着样本数量的不足，局部结构信息比全局结构信息更为重要，无法充分地利用样本信息，导致故障诊断效果下降，识别准确率低。

为了提高本研究所提LCMMDP算法的泛化能力，将不同算法在不同转速（分别为2800，3000，3200 r/min）下降维得到的低维特征子集输入KNCNCM分类器中进行故障识别，得到的平均识别准确率如图7所示。由图7可以看出，在不同转速下，LCMMDP降维方法的平均识别准确率都明显优于其他5种降维方法，表明它具有良好的适用性和更高的故障识别精度。

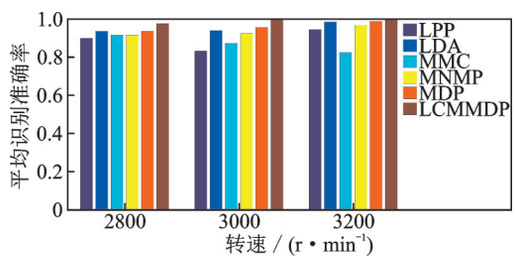


图7 不同算法在不同转速下的平均识别准确率

Fig. 7 The average recognition accuracy of different algorithms at different speeds

5.2.4 LCMMDP算法的抗干扰能力分析

为了分析LCMMDP的抗干扰能力，将系数γ=0.1，0.2，0.3，0.4的rand随机干扰噪声加入到原始故障集中[19-20]，经上述6种算法降维后得到低维特征集，然后输入KNCNCM分类器进行故障识别，得到的识别准确率如图8所示。由图可知，随着干扰系数的增加，6种算法的准确率都有所降低，但LCMMDP降低的速率较慢，且准确率都明显高于其他算法。因此可以看出LCMMDP的抗干扰能力强，相应的鲁棒性较好。

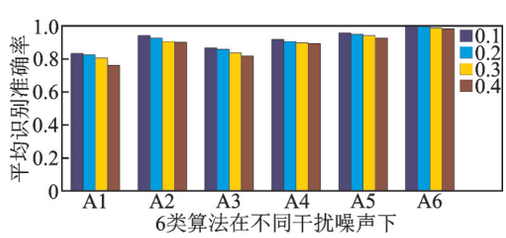


图8 不同随机噪声干扰下各类算法的平均识别准确率

Fig. 8 The average recognition accuracy of various algorithms under different random noise interferences

5.2.5 KNCNCM分类器的性能分析

为了验证本文所提KNCNCM分类器的鲁棒性和稳定性，向测试故障集中融入系数为a=0.1，0.2，0.3，0.4的随机干扰噪声，将经LCMMDP降维后的低维敏感特征集输入KNCNCM，LMKNCN，KNN分类器中进行故障模式识别，得到的平均识别率如表5所示。

由表5中可以看出，LMKNCN的识别率高于KNN，原因是KNN的分类性能容易受到噪声干扰

表 5　不同分类器的抗干扰能力对比

Tab. 5　The anti-interference comparison of different classifiers

分类器	在不同干扰系数下的平均识别率/ %			
	a=0.1	a=0.2	a=0.3	a=0.4
KNCNCM	98.8	97.2	93.6	90.0
LMKNCN	98.8	94.0	89.2	84.0
KNN	100	92.0	86.4	83.6

和离群点的影响，而LMKNCN利用每类训练样本集里未分类样本的几个近邻的局部均值信息，一定程度上克服了离群点对分类性能的影响；随着噪声的增多，KNCNCM的诊断结果变化较小，平均识别率都高于其他两种分类器，表明KNCNCM相较于LMKNCN和KNN对噪声不敏感，具有优异的稳定性和识别能力。

6　结　论

为使提取的旋转机械故障特征有利于实施故障数据集分类，本研究提出一种基于局部质心均值最小距离鉴别投影(LCMMDP)降维方法和基于局部均值与类均值的k-近质心近邻分类方法(KNCNCM)相结合的旋转机械故障诊断方法。分别通过一个双转子系统的振动信号集合和仿真数据集进行验证，实验结果表明：

(1) LCMMDP相比较于MDP，MNMP，LDA，LPP，MMC等降维方法，可提取出可分性更高的低维空间故障特征集，在进行故障模式识别时具有一定的优势。

(2) KNCNCM分类方法既利用未分类样本在每类里的近质心近邻的局部信息，又利用了类均值的整体知识，克服了数据离群点对分类性能的影响，而且一定程度上避免了噪声的干扰，具有一定的稳定性和准确性。

(3) LCMMDP与KNCNCM相结合的维数约简故障诊断模式能够有效地对高维转子故障数据集进行维数约简和故障分类，为旋转机械智能故障诊断提供了一种解决方案。

参考文献：

[1] Wang Z Y, Lu C, Zhou B. Fault diagnosis for rotary machinery with selective ensemble neural networks[J]. Mechanical Systems and Signal Processing, 2018, 113: 112-130.

[2] SU Z, Tang B, MA J, et al. Fault diagnosis method based on incremental enhanced supervised locally linear embedding and adaptive nearest neighbor classifier[J]. Measurement, 2014, 48(1): 136-148.

[3] Turk M, Pentland A. Eigenfaces for recognition[J]. Journal of Cognitive Neuroscience, 1991, 3(1):71-86.

[4] Martinez A M, Kak A C. PCA versus LDA[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(2): 228-233.

[5] He X. Locality preserving projections[J]. Advances in Neural Information Processing System, 2003, 16(1): 186-197.

[6] ROWEIS S T, SAUL L K. Nonlinear dimensionality reduction by locally linear embedding[J]. Science, 2000, 290(5500): 2323-2326.

[7] JIANG Quansheng, JIA Minping, HU Jianzhong, et al. Modified Laplacian eigenmap method for fault diagnosis[J]. Chinese Journal of Mechanical Engineering, 2008, 21(3): 90-93.

[8] Sugiyama M. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis[J]. Journal of Machine Learning Research, 2007, 8(1): 1027-1061.

[9] Yan S C, Xu D, Zhang B Y, et al. Graph embedding and extensions: A general framework for dimensionality reduction[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(1): 40-51.

[10] Yu W, Teng X, Liu C. Face recognition using discriminant locality preserving projections[J]. Image and Vision Computing, 2006, 24(3): 239-248.

[11] 黄　璞，唐振民. 最小距离鉴别投影及其在人脸识别中的应用[J]. 中国图象图形学报，2013，18(02)：201-206.
Huang Pu, Tang Zhenmin. Minimum-distance discriminant projection and its application to face recognition[J]. Journal of Image and Graphics, 2013, 18(02): 201-206.

[12] 黄　璞，唐振民. 鉴别的局部中值保持投影及其在人脸识别中的应用[J]. 计算机辅助设计与图形学学报，2012，24(11)：1420-1425.
Huang Pu, Tang Zhenmin. Discriminant local median preserving projections with its application to face recognition[J]. Journal of Computer-Aided Design & Computer Graphics, 2012, 24(11): 1420-1425.

[13] Gou J, Yi Z, Du L, et al. A local mean-based k-nearest centroid neighbor classifier[J]. The Computer Journal, 2012, 55(9): 1058-1071.

[14] 曾　勇，杨煜普，赵　亮. 基于局部均值与类均值的近邻分类[J]. 控制与决策，2009，24(04)：547-550.
Zeng Yong, Yang Yupu, Zhao Liang. Nearest neighbor classification based on local mean and class mean[J]. Control and Decision, 2009, 24(04): 547-550.

[15] 张晓涛，唐力伟，王　平，等. 基于半监督PCA-LPP

流形学习算法的故障降维辨识[J]. 中南大学学报(自然科学版), 2016, 47(05): 1559-1564.
ZHANG Xiaotao, TANG Liwei, WANG Ping, et al. Fault identification and dimensionality reduction method based on semi-supervised PCA-LPP manifold learning algorithm[J]. Journal of Central South University (Science and Technology), 2016, 47(05): 1559-1564.

[16] BLAKE C L, MERZ C J, 1998, UCI repository of machine learning databases[EB/OL]. [2019-01-12] University of California. Available online at: http://archive.Ics.Uci. edu/ml/.

[17] 苏祖强, 汤宝平, 姚金宝. 基于敏感特征选择与流形学习维数约简的故障诊断[J]. 振动与冲击, 2014, 33(03): 70-75.
SU Zu-qiang, Tang Bao-ping, Yao Jin-bao. Fault diagnosis based on sensitive feature selection and manifold learning dimension reduction[J]. Journal of Vibration and Shock, 2014, 33(03): 70-75.

[18] 李学军,李 平,蒋玲莉. 类均值核主元分析法及在故障诊断中的应用[J]. 机械工程学报, 2014, 50(03): 123-129.
LI Xue-Jun, LI Ping, Jiang Ling-Li. Class mean kernel principal component analysis and its application in fault diagnosis [J]. Journal of Mechanical Engineering, 2014, 50(03): 123-129.

[19] Chen F, Tang B, Chen R. A novel fault diagnosis model for gearbox based on wavelet support vector machine with immune genetic algorithm [J]. Measurement, 2013, 46(1): 220-232.

[20] 赵孝礼, 赵荣珍. 全局与局部判别信息融合的转子故障数据集降维方法研究[J]. 自动化学报, 2017, 43(04): 560-567.
Zhao Xiao-li, Zhao Rong-zhen. A method of dimension reduction of rotor faults data set based on fusion of global and local discriminant information[J]. Acta Automatica Sinica, 2017, 43(04): 560-567.

Dimensional reduction analysis of rotating machinery fault data based on local centroid mean minimum-distance discriminant projection

SHI Ming-kuan, ZHAO Rong-zhen

(School of Mechanical and Electronical Engineering, Lanzhou University of Technology, Lanzhou 730050, China)

Abstract: Aiming at the problem of classification difficulty caused by the strong nonlinearity and the high dimensionality of fault dataset of rotating machinery, a fault dataset dimension reduction algorithm local centroid mean minimum-distance discriminant projection (LCMMDP) is proposed. The algorithm can maintain the local geometric structure information of the sample while considering the cohesion and separation of the sample, reflecting the close relationship between the sample and the local centroid mean. The hybrid characteristics of rotor vibration signals are extracted from multiple angles, the original high-dimensional feature sets are constructed, and low-dimensional sensitive feature subsets are extracted by LCMMDP. The improved k-nearest centroid neighbor classification based on local mean and class mean is used (KNCNCM) for fault pattern recognition. The proposed method integrates the advantages of LCMMDP in dimension reduction and KNCNCM in pattern recognition and provides higher fault identification accuracy. The validity of the proposed method is verified by the instance of the fault diagnosis of a double-span rotor system dataset and simulation dataset.

Key words: fault diagnosis; dimension reduction; local centroid mean; classifier; pattern recognition

作者简介: 石明宽 (1993-),男,硕士研究生。电话:18809490031;E-mail: 1937787272@qq.com
通讯作者: 赵荣珍 (1960-),女,教授,博士生导师。电话:13619349619;E-mail: zhaorongzhen@lut.cn