DOI:10.19651/j.cnki.emt.2314111

基于多信息融合的自适应局部线性嵌入算法*

刘庆强 魏朝阳
(东北石油大学电气信息工程学院 大庆 163318)

摘 要:局部线性嵌入算法 LLE 的降维性能与挖掘的流形结构密切相关,但 LLE 挖掘的流形结构单一,并且对邻域参数选取敏感,无法提取全面的流形局部结构,限制了 LLE 的降维性能。为此,本文提出基于多信息融合的自适应局部线性嵌入算法 MIF-ALLE。MIF-ALLE 首先利用切空间近似判据自适应选择邻域参数,获取更准确的局部邻域;然后,将局部邻域中蕴含的切空间角度信息与局部线性信息相融合,挖掘更全面的流形局部结构,降低局部低维嵌入的偏差;最后,在公开轴承数据集以及实验室提取的轴承数据集上进行实验验证。实验结果表明:MIF-ALLE 可以挖掘更全面的流形结构,提取更显著的特征,轴承故障诊断准确率最高可达 100%。

关键词:局部线性嵌入算法;流形结构;自适应邻域;轴承故障诊断
中图分类号:TH133.3 **文献标识码**:A **国家标准学科分类代码**:510

Adaptive local linear embedding algorithm based on multiple information fusion

Liu Qingqiang Wei Zhaoyang
(School of Electrical Information Engineering, Northeast Petroleum University, Daqing 163318, China)

Abstract: The dimensionality reduction performance of Local Linear Embedding algorithm LLE is closely related to the manifold structure mined. However, the manifold structure mined by LLE is singular and sensitive to the selection of neighborhood parameters, making it difficult to extract a comprehensive local structure of the manifold, which limits its dimensionality reduction performance. Therefore, this article proposes an adaptive local linear embedding algorithm based on multiple information fusion MIF-ALLE. MIF-ALLE firstly uses tangent space approximation criterion to adaptively select neighborhood parameters to obtain more accurate local neighborhood; Then, the angle information of Tangent space contained in the local neighborhood is fused with the local linear information to mine a more comprehensive local structure of the manifold and reduce the deviation of local low dimensional embedding; Finally, the experimental verification is carried out on the bearing data set published and the bearing data set extracted from the laboratory. The experimental results show that MIF-ALLE can mine more comprehensive manifold structures, extract more significant features, and achieve bearing fault diagnosis accuracy of up to 100%.

Keywords: local linear embedding; adaptive neighborhood; manifold structure; bearing fault-diagnosis

0 引 言

目前,轴承数据日益增多,并呈现出高维度和稀疏性[1-2]。挖掘高维稀疏数据中的潜在低维结构是数据分析过程(如特征提取、模式分类以及故障诊断)的必要预处理步骤。流形学习算法的提出为这类问题提供了有效的解决方案[3]。

流形学习是一种高效的数据降维技术,能够揭示高维数据的潜在低维结构,同时最大限度地保留局部特征。作为经典的流形学习算法,局部线性嵌入算法(local linear embedding, LLE)[4]由于其计算复杂度低和可扩展性强的特性在图像处理、语音识别、故障诊断等领域被广泛应用[5]。局部线性嵌入算法的框架由以下 3 个部分构成:首先,构造局部近邻域;其次,在每个样本的局部邻域内进行线性逼近;最后,最小化全局误差函数以获得高维流形的低维表示。

LLE 算法主要存在两个问题[6]。第 1 个问题是 LLE 算法对邻域参数的选取非常敏感,不同的近邻参数对降维

收稿日期:2023-07-14
* 基金项目:海南省自然科学基金(623MS071)项目资助

结果有很大影响。为了优化此问题,基于核方法、聚类思想以及信息熵概念的自适应邻域算法被提出。文献[7]使用核密度估计每个样本的邻域大小,文献[8]通过对核矩阵进行特征分解自适应选择邻域;文献[9-10]利用聚类算法将数据划分成不同的簇群,然后在每个簇群内选择邻近样本构建局部邻域;文献[11-12]基于信息熵度量邻域集合的不确定性,通过最小化信息熵确定每个样本的邻域大小。

第 2 个问题是流形局部信息的挖掘,流形局部内蕴丰富的几何信息,而 LLE 算法只考虑了流形局部的线性结构。为了优化此问题,基于稀疏表示、正则化准则以及权重融合思想的改进 LLE 算法被提出。文献[13-14]基于稀疏表示和正则化准则,将局部线性结构的嵌入扩展到多尺度空间的同时加入正则化项优化目标函数;文献[15-16]根据局部密度等信息计算新的权重,并将其与局部线性权重融合,挖掘更全面的流形局部结构。但是,上述算法没有考虑切空间中蕴含的丰富流形局部结构信息。

基于上述讨论,本文提出一种基于多信息融合的自适应局部线性嵌入算法(adaptive local linear embedding algorithm based on multiple information fusion, MIF-ALLE),主要贡献如下:

1)提出一种基于切空间近似判据的自适应邻域策略,解决了 LEE 算法邻域参数选取敏感问题。

2)将局部邻域中蕴含的局部切空间角度信息与局部线性信息相融合,挖掘了更全面的流形局部结构,解决了 LEE 算法挖掘流形结构单一问题。

3)在两个轴承数据集上进行实验验证,实验结果表明了 MIF-ALLE 算法在挖掘流形结构、特征提取方面的优异性能。

1 相关工作

1.1 局部线性嵌入算法

局部线性嵌入算法是一种经典的非线性降维算法,它能够从高维流形中学习样本的内在拓扑,并使映射到低维空间的样本保持这种拓扑关系。

给定数据集 $\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N)\in R^{D\times N}$,其中 D 表示数据集的原始维数,N 表示数据集的样本数量。通过 LLE 算法降维后得到 $\boldsymbol{Y}=(\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_N)\in R^{d\times N}$,其中 $d<<D$。局部线性嵌入算法具体流程如下:

步骤 1)使用 k 近邻算法计算出每个样本点 $\boldsymbol{x}_i$ 的 k 个近邻点;

步骤 2)通过最小化代价函数,计算权重系数矩阵,即求解以下有约束优化问题:

$$\underset{\boldsymbol{W}}{\operatorname{argmin}}\parallel \boldsymbol{x}_i-\sum_{j=1}^{k}w_{ij}\boldsymbol{x}_j\parallel_2^2,\ \text{s.t.}\sum_{j=1}^{k}w_{ij}=1 \tag{1}$$

其中,$\boldsymbol{x}_j(j=1,2,\cdots,k)$ 是 $\boldsymbol{x}_i$ 的 k 个近邻点,w_{ij} 是 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 之间的线性重构权值;

步骤 3)通过最小化损失函数 $J(\boldsymbol{y})$ 来获得最优的低维嵌入结果。

$$J(\boldsymbol{y})=\sum_{i=1}^{N}\parallel \boldsymbol{y}_i-\sum_{j=1}^{k}w_{ij}\boldsymbol{y}_j\parallel_2^2$$
$$\text{s.t.}\sum_{i=1}^{N}\boldsymbol{y}_i=0\ \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{y}_i\boldsymbol{y}_i^{\mathrm{T}}=I \tag{2}$$

其中,$\boldsymbol{y}_i$ 与 $\boldsymbol{y}_j$ 分别是 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 降维后的结果,最小化损失函数可以展开如下:

$$\sum_{i=1}^{N}\parallel \boldsymbol{y}_i-\sum_{j=1}^{k}w_{ij}\boldsymbol{y}_j\parallel_2^2=\sum_{i=1}^{N}\parallel \boldsymbol{YI}_i-\boldsymbol{YW}_i\parallel_2^2=$$
$$\operatorname{tr}(\boldsymbol{Y}(\boldsymbol{I}-\boldsymbol{W})(\boldsymbol{I}-\boldsymbol{W})^{\mathrm{T}}\boldsymbol{Y}^{\mathrm{T}})=\operatorname{tr}(\boldsymbol{YMY}^{\mathrm{T}}) \tag{3}$$

其中,$\boldsymbol{I}_i\in R^{d\times 1}$ 为单位向量,$\boldsymbol{I}\in R^{d\times d}$ 为单位矩阵,$\boldsymbol{W}_i\in R^{N\times 1}$ 为样本点 $\boldsymbol{x}_i$ 的重构向量,$\boldsymbol{W}\in R^{N\times N}$ 为数据集 $\boldsymbol{X}$ 的重构矩阵,$\boldsymbol{M}=(\boldsymbol{I}-\boldsymbol{W})(\boldsymbol{I}-\boldsymbol{W})^{\mathrm{T}}$。为了最小化损失函数,$\boldsymbol{Y}$ 取矩阵 $\boldsymbol{M}$ 的前 d 个最小非零特征值对应的特征向量。

1.2 切空间

切空间是用来描述流形上样本点处局部性质的概念。高维流形中的局部切空间不能直接计算,一种常见的求解方法是在局部邻域上执行主成分分析(principal component analysis, PCA)[17]。然而,在固定局部邻域参数的情况下,通过 PCA 求得的切空间会受到离群点的影响,从而与真实的切空间产生偏差,如图 1 所示。

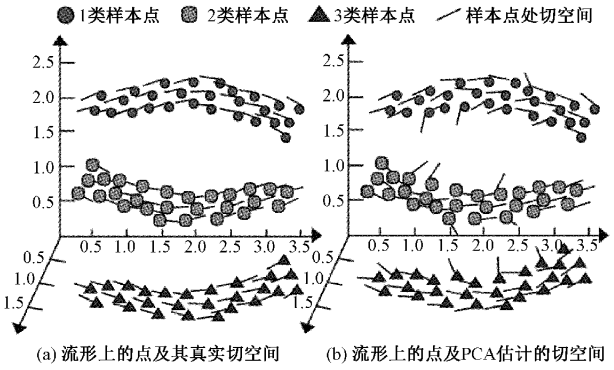

图 1 真实切空间与 PCA 估计的切空间对比

为解决此问题,本文提出自适应邻域策略,此策略选择出的自适应邻域能够帮助每个样本点获取更逼近真实流形的切空间。

1.3 度量方式

传统的欧式度量方法并不适用于高维稀疏数据,余弦相似度更适用于在高维稀疏数据中寻找近邻点[18],余弦相似度定义如下:

$$c(\boldsymbol{x}_i,\boldsymbol{x}_j)=\frac{\boldsymbol{x}_i^{\mathrm{T}}\cdot\boldsymbol{x}_j}{\mid \boldsymbol{x}_i\mid\mid \boldsymbol{x}_j\mid} \tag{4}$$

其中,余弦相似度 $c(\boldsymbol{x}_i,\boldsymbol{x}_j)\in[0,1]$,其值接近于 1,两个样本点相似性就越高。

2 基于多信息融合的自适应局部线性嵌入算法

2.1 基于切空间近似判据的自适应邻域策略

由切空间近似判据可知,采样点处的切空间可以被视

为该点的局部线性逼近。为消除离群点的影响，每个采样点选择的局部邻域构成的流形应该是相对平坦的，这样通过对局部邻域的最优线性拟合确定的线性子空间才能更准确的逼近切线空间。

如图 2 所示，在流形 M 上，点 $\boldsymbol{x}$ 处的曲率很小，取较大或较小的局部邻域参数构成的流形都较为平坦；点 $\boldsymbol{x}'$ 处的曲率较大，只有较小的局部邻域参数才能使构成的流形相对平坦。

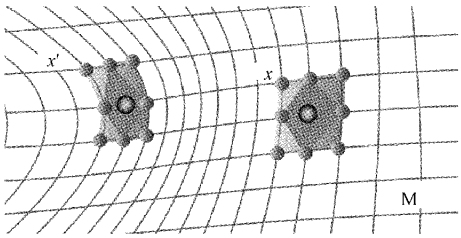

图 2　流形上不同曲率处的邻域选取

设定 $\boldsymbol{x}=f(\boldsymbol{\tau})$，$f$ 是切线空间到 $\boldsymbol{x}$ 处的邻域空间的映射函数，$\boldsymbol{\tau}$ 是 $\boldsymbol{x}$ 在切空间中对应的投影点。假设 f 足够光滑，在给定的 $\boldsymbol{\tau}$ 处做泰勒展开：

$$\tilde{\boldsymbol{x}}=\boldsymbol{x}+Q_f(\boldsymbol{\tau})\cdot(\tilde{\boldsymbol{\tau}}-\boldsymbol{\tau})+\frac{1}{2}\|\boldsymbol{H}_{\tau}\|\,\|\tilde{\boldsymbol{\tau}}-\boldsymbol{\tau}\|^2 \tag{5}$$

其中，$Q_f(\boldsymbol{\tau})\in R^{D\times d}$ 是 f 在 $\boldsymbol{\tau}$ 处的雅可比矩阵，以其列向量为基底张成了 f 在 $\boldsymbol{\tau}$ 处的切空间。$\boldsymbol{H}_{\tau}$ 是 f 在 $\boldsymbol{\tau}$ 处的海森矩阵，它刻画了流形的局部曲率。为了保证在 $\boldsymbol{x}_i$ 处所选邻域的局部流形是相对平坦的，构造以下不等式：

$$\sum_{j=1}^{K_i}\|\boldsymbol{x}_{ij}-\boldsymbol{x}_i-Q_i(\boldsymbol{\tau}_{ij}-\boldsymbol{\tau}_i)\|_2\leqslant\beta\sum_{j=1}^{K_i}\|Q_i(\boldsymbol{\tau}_{ij}-\boldsymbol{\tau}_i)\|_2 \tag{6}$$

其中，$\boldsymbol{x}_{ij}$ 是 $\boldsymbol{x}_i$ 的近邻点，$\boldsymbol{\tau}_{ij}$ 是 τ_i 的近邻点，K_i 是 $\boldsymbol{x}_i$ 的邻域参数，β 用于确定局部线性结构，它意味着二阶项要比一阶项小得多，通过考虑泰勒展开的平均值形式得到以下公式：

$$\sum_{j=1}^{K_i}\|\boldsymbol{x}_{ij}-\overline{\boldsymbol{x}_i}-Q_i(\boldsymbol{\tau}_{ij}-\overline{\boldsymbol{\tau}_i})\|_2\leqslant\beta\sum_{j=1}^{K_i}\|Q_i(\boldsymbol{\tau}_{ij}-\overline{\boldsymbol{\tau}_i})\|_2 \tag{7}$$

其中，$\bar{\boldsymbol{x}}_i\in R^{D\times 1}$ 是由 $\boldsymbol{x}_i$ 的平均值构成的向量，$\overline{\boldsymbol{\tau}}_i\in R^{D\times 1}$ 是由 $\boldsymbol{\tau}_i$ 的平均值构成的向量，为方便计算，考虑到公式的矩阵形式如下：

$$\|\boldsymbol{X}_i-(\overline{\boldsymbol{x}_i}\boldsymbol{e}^{\mathrm{T}}+\boldsymbol{Q}_i\boldsymbol{\Theta}_i)\|_F\leqslant\beta\|\boldsymbol{\Theta}_i\|_F \tag{8}$$

其中，$\boldsymbol{X}_i$ 是由 $\boldsymbol{x}_i$ 的 K_i 个近邻点构成的矩阵，$\boldsymbol{e}$ 为单位向量，$\boldsymbol{\Theta}_i$ 是由 $\boldsymbol{\theta}_{ij}$ 构成的矩阵，$\boldsymbol{\theta}_{ij}=Q_i^{\mathrm{T}}(\boldsymbol{x}_{ij}-\overline{\boldsymbol{x}_i})$。在实际计算中，用 $\boldsymbol{X}_i-\overline{\boldsymbol{x}_i}\boldsymbol{e}^{\mathrm{T}}$ 的奇异值 $\sigma_j^i(0\leqslant j\leqslant K_i)$ 代表式(8)进行计算。

$$\|\boldsymbol{\Theta}_i\|_F=\sqrt{\sum_{j\leqslant n}(\sigma_j^i)^2} \tag{9}$$

$$\|\boldsymbol{X}_i-(\overline{\boldsymbol{x}_i}\boldsymbol{e}^{\mathrm{T}}+\boldsymbol{Q}_i\boldsymbol{\Theta}_i)\|_F=\sqrt{\sum_{n<j\leqslant K_i}(\sigma_j^i)^2} \tag{10}$$

使用式(11)进行自适应邻域选择：

$$\frac{\sqrt{\sum_{n<j\leqslant K_i}(\sigma_j^i)^2}}{\sqrt{\sum_{j\leqslant n}(\sigma_j^i)^2+\sum_{n<j\leqslant K_i}(\sigma_j^i)^2}}=\lambda \tag{11}$$

式中：n 值一般取 2，代表了切空间标准正交基数量为 2。定义逼近系数 λ，它的值越接近于 0，说明由 $\boldsymbol{x}_i$ 的 K_i 个近邻点构成的流形越平坦，由其求得的切空间越逼近于原始流形。

利用式(11)作为自适应邻域准则，提出了一种自适应邻域策略如算法 1 所示。

算法 1 基于切空间近似判据的自适应邻域策略

输入：原始数据 $\boldsymbol{X}$，切空间标准正交基数量 $n=2$，最大邻域阈值 $k_{\max}$，最小邻域阈值 $k_{\min}$

输出：自适应邻域 $\boldsymbol{K}$

1：for i=1 to N：

2：for $k=k_{\min}$ to $k_{\max}$：

3：　通过式(4)寻找点 $\boldsymbol{x}_i$ 的 k 近邻

4：　通过式(11)求得点 $\boldsymbol{x}_i$ 的 k 近邻逼近系数 λ_{ik}

5：取最小的逼近系数 λ_{ik} 对应的 k 值，令 $\boldsymbol{K}_i=k$

6：得到自适应邻域 $\boldsymbol{K}$

2.2　局部切空间角度信息

通过本文提出的自适应邻域策略求得每个样本点最逼近原始流形的切空间后，考虑样本点 x_i 的邻域点 x_{ij} 与 x_i 处切空间之间的角度信息。

如图 3 所示，定义 $\boldsymbol{e}_{ij}=\boldsymbol{x}_{ij}-\boldsymbol{x}_i$，$\boldsymbol{e}_{ij}$ 到切空间的投影 $\boldsymbol{pro}_{ij}$，$\boldsymbol{e}_{ij}$ 求与 $\boldsymbol{x}_i$ 处切空间之间的夹角 w_{ij}，由于 $\cos\boldsymbol{w}_{ij}\in[0,1]$，将 $\cos\boldsymbol{w}_{ij}$ 称为点 $\boldsymbol{x}_i$ 与其邻域点 $\boldsymbol{x}_{ij}$ 之间的切空间偏移量。

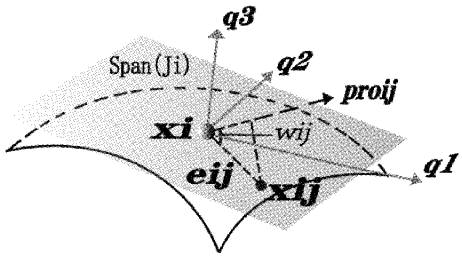

图 3　切空间偏移量

点 $\boldsymbol{x}_i$ 与其邻域点 $\boldsymbol{x}_{ij}$ 之间的切空间偏移量公式如下：

$$\cos\boldsymbol{w}_{ij}=\frac{\boldsymbol{pro}_{ij}\cdot\boldsymbol{e}_{ij}}{|\boldsymbol{pro}_{ij}||\boldsymbol{e}_{ij}|} \tag{12}$$

切空间偏移量越接近于 0，说明 $\boldsymbol{x}_{ij}$ 越偏离 $\boldsymbol{x}_i$ 处的切平面，应减少其权重；切空间偏移量越接近于 1，说明 $\boldsymbol{x}_{ij}$ 越接近 $\boldsymbol{x}_i$ 处的切平面，应增加其权重。

2.3　多信息融合过程

为方便与线性权重相融合，将切空间偏移量归一化如下：

$$\overline{\boldsymbol{W}_{ij}}=\frac{\cos\boldsymbol{w}_{ij}}{\sum_{j=1}^{K_i}\cos\boldsymbol{w}_{ij}} \tag{13}$$

利用以下公式将线性重构权重 $\boldsymbol{W}$ 与切空间偏移量 $\overline{\boldsymbol{W}}$ 相结合得到多信息融合权重 $\boldsymbol{W}^*$，具体过程如图 4 所示。

$$\boldsymbol{W}_{ij}^*=\alpha\boldsymbol{W}_{ij}+(1-\alpha)\overline{\boldsymbol{W}_{ij}} \tag{14}$$

其中，α 是权衡系数，用于权衡线性重构权重和切空间偏移量所占的比例，将多信息融合权重 $\boldsymbol{W}^*$ 应用于求解低维嵌入结果如式(15)所示。

$$\begin{aligned}&\underset{\boldsymbol{y}}{\operatorname{argmin}}\sum_{i=1}^{N}\|\boldsymbol{y}_i-\sum_{j=1}^{K_i}w_{ij}^*\boldsymbol{y}_j\|_2^2=\\&\underset{\boldsymbol{Y}}{\operatorname{argmin}}\sum_{i=1}^{N}\|\boldsymbol{YI}_i-\boldsymbol{YW}_i^*\|_2^2=\\&\underset{\boldsymbol{Y}}{\operatorname{argmin}}\operatorname{tr}(\boldsymbol{Y}(\boldsymbol{I}-\boldsymbol{W}^*)(\boldsymbol{I}-\boldsymbol{W}^*)^{\mathrm{T}}\boldsymbol{Y}^{\mathrm{T}})=\underset{\boldsymbol{Y}}{\operatorname{argmin}}\operatorname{tr}(\boldsymbol{YM}^*\boldsymbol{Y}^{\mathrm{T}})\end{aligned} \tag{15}$$

其中，$\boldsymbol{W}^*$ 为多信息融合权重，$\boldsymbol{M}^*=(\boldsymbol{I}-\boldsymbol{W}^*)(\boldsymbol{I}-\boldsymbol{W}^*)^{\mathrm{T}}$。

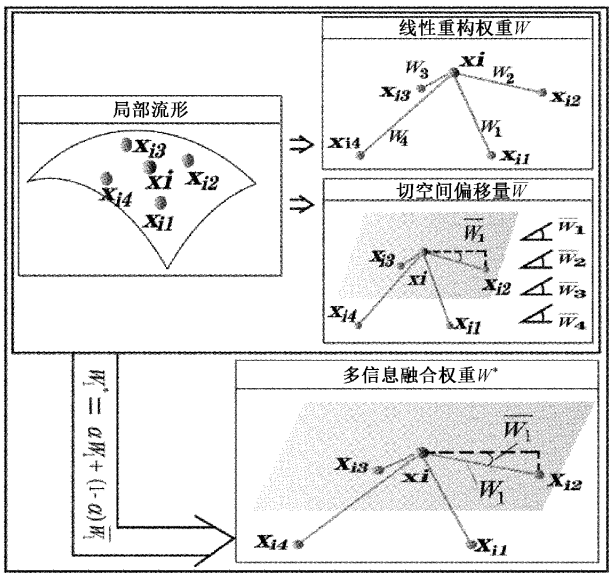

图 4 多信息融合过程

2.4 MIF-ALLE 算法

MIF-ALLE 算法的具体描述如算法 2 所示。

算法 2 基于多信息融合的自适应局部线性嵌入算法

输入：原始数据 $\boldsymbol{X}$，目标维度 $d=3$，标准正交基数量 $n=2$，最大邻域阈值 $k_{\max}$，最小邻域阈值 $k_{\min}$，权衡系数 α

输出低维嵌入 $\boldsymbol{Y}$

1：由算法 1 得到自适应邻域 $\boldsymbol{K}$

2：由式(2)得到局部线性重构权重系数 $\boldsymbol{W}$

3：由式(13)求得切空间偏移量 $\overline{\boldsymbol{W}}$

4：由式(14)求得多信息融合权重 $\boldsymbol{W}^*$

5：由式(15)求得低维嵌入 $\boldsymbol{Y}$

3 实验结果与分析

将 MIF-ALLE 算法分别在凯斯西储大学(Case Western Reserve University，CWRU)滚动轴承数据中心的公开数据集和东北石油大学(Northeast Petroleum University，NEPU)轴承故障仿真平台采集到的数据集上进行实验，两个数据集的部分参数如表 1。从可视化和聚类效果的角度出发，与其它流形学习算法进行比较，并对算法参数进行分析，证明了本文提出算法的优越性。

表 1 两个数据集的部分参数

数据集	采样频率/kHz	类别数	样本数量	样本维度
CWRU	12	4	400	1 024
NEPU	10	4	400	1 024

3.1 实验数据集

数据集 1：数据集 1 采集于凯斯西储大学(CWRU)滚动轴承数据中心，测试平台如图 5 所示。从驱动端轴承采集信号，采用电火花加工形成轴承故障，原始振动数据在电机转速为 1 797 RPM 的情况下，由加速度计以 12 kHz 的采样频率从电机驱动机械系统获得。测试平台模拟了 4 种不同的数据类型，即轴承正常状态、球故障、内圈故障和外圈故障。每种数据的样本数量为 100，特征维度为 1 024。

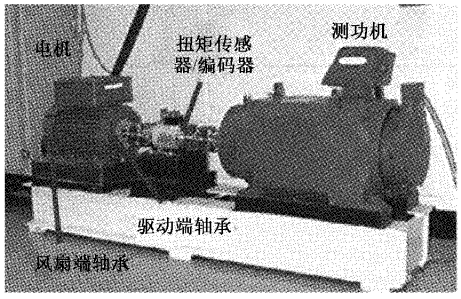

图 5 CWRU 测试平台图

数据集 2：数据集 2 采集于东北石油大学(NEPU)轴承故障测试平台，测试平台如图 6 所示。原始振动数据在电机转速为 1 400 RPM 的情况下，通过加速度计以 10 kHz 的采样频率在该平台上获得。测试平台模拟了 4 种不同的数据类型，即轴承正常状态、球故障、内圈故障和外圈故障。每种数据的样本数量为 100，特征维度为 1 024。

3.2 可视化实验

为了定性分析算法的特征提取能力，本节从降维可视化效果的角度出发，将所提出的 MIF-ALLE 算法分别应用于上述两个轴承数据集，并与下述经典降维算法进行比较。

1)主成分分析(principal component analysis，PCA)；

2)局部切空间排列(local tangent space alignment，LTSA)；

3)拉普拉斯映射(Laplacian eigenmaps，LE)；

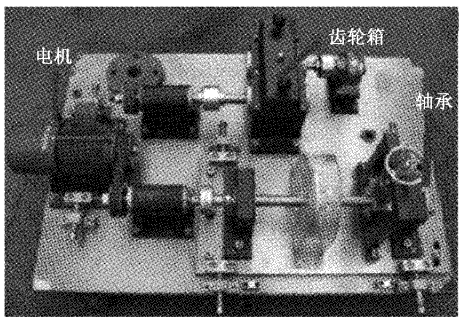

图 6　NEPU 测试平台

4)局部线性嵌入(local linear embedding，LLE)；

5)半监督局部线性嵌入(semi-supervised local li-near embedding，SSLLE)。

此外，本文选择展示非自适应局部降维算法可视化效果最好的结果：LTSA(k=12,10)、LE(k=11,8)、LLE(k=11,7)、SSLLE(k=15,8)。

图 7 和 8 分别对应 6 个算法在数据集 1 以及数据集 2 上的可视化实验结果，其中红色点代表正常状态数据、绿色点代表内圈故障数据、蓝色点代表球故障数据、黑色代表外圈故障数据。分析图 7 和 8 可知，LTSA 算法可视化结果的类内样本紧凑度以及类间样本分离度都相对较差；PCA、LE、LLE、SSLLE 算法可视化结果的类内样本紧凑度较好，

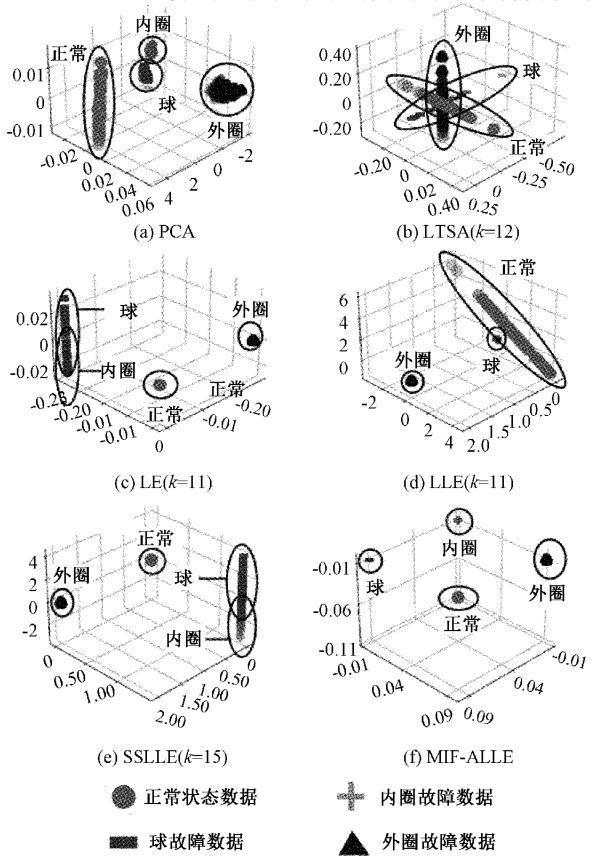

图 7　数据集 1 上的可视化实验结果

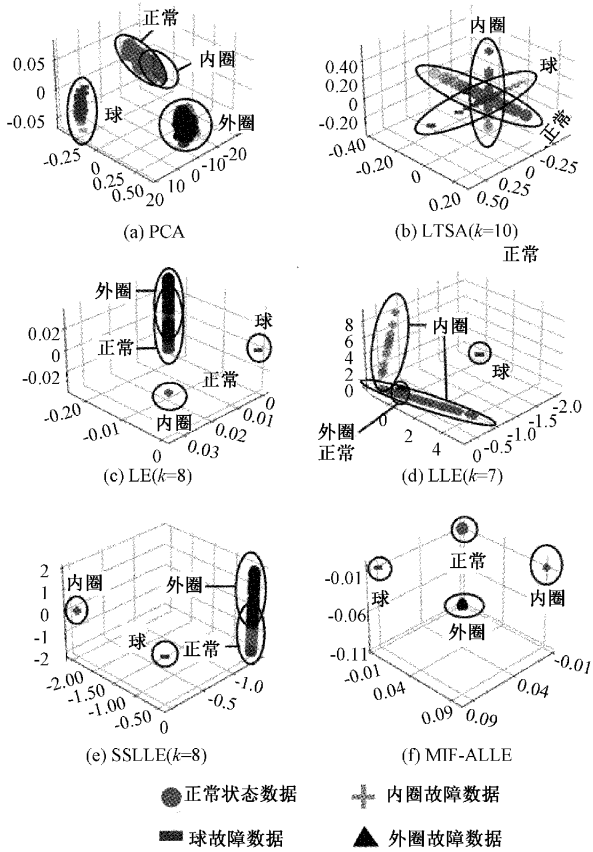

图 8　数据集 2 上的可视化实验结果

但类间样本的分离度较差，即不同类样本之间存在重叠以及覆盖的情况；与其它算法相比，本文提出的 MIF-ALLE 算法不仅能够清晰的分离不同类型的数据，而且还能增强同类型数据的紧凑程度。

这说明了本文提出的自适应邻域策略可以剔除不同类样本，增强类间分离度；局部切角信息与局部线性信息相融合可以挖掘出更全面的流形结构，增强类内紧凑度；MIF-ALLE 算法具有良好的鲁棒性和可视化效果。

3.3　聚类实验

为了定量分析算法的聚类性能，本节将利用聚类精度(cluster accuracy，ACC)和标准化互信息(normalized mutual information，NMI)两个指标进一步定量评估 MIF-ALLE 算法的降维效果。ACC 和 NMI 是目前机器学习领域比较流行的评价指标，它们可以有效地衡量聚类结果与实际类别标签之间的匹配程度。当聚类结果与真实类别标签完全匹配时，ACC 和 NMI 均为最大值 1。

为减少随机因素的影响，对两个数据集降维后的结果分别进行 10 次支持向量机(support vector machines，SVM)聚类实验，每次实验随机选取各类型样本 80 个作为训练集，20 个作为测试集，实验结果如表 2 所示。通过分析可知，聚类实验结果与上述可视化实验结果遥相呼应，MIF-ALLE 算法的 ACC 以及 NMI 指标都高于其他算法，

体现了此算法在处理轴承数据时具有优异的特征提取能力。

表 2 聚类实验结果

算法	数据集 1		数据集 2	
	ACC	NMI	ACC	NMI
PCA	0.859 4	0.789 1	0.900 0	0.823 5
LTSA	0.545 0	0.381 6	0.534 4	0.352 9
LE	0.853 1	0.791 0	0.865 6	0.800 2
LLE	0.859 4	0.793 7	0.848 4	0.733 9
SSLLLE	0.909 4	0.843 5	0.942 27	0.882 2
MIF-ALLE	1.000 0	1.000 0	1.000 0	1.000 0

3.4 参数评估

本节将探寻权衡参数 α 对 MIF-ALLE 算法的影响,首先,在权衡参数 $\alpha \in [0,1]$ 的情况下对两个数据集进行可视化实验,可视化实验结果如图 9、10 所示。

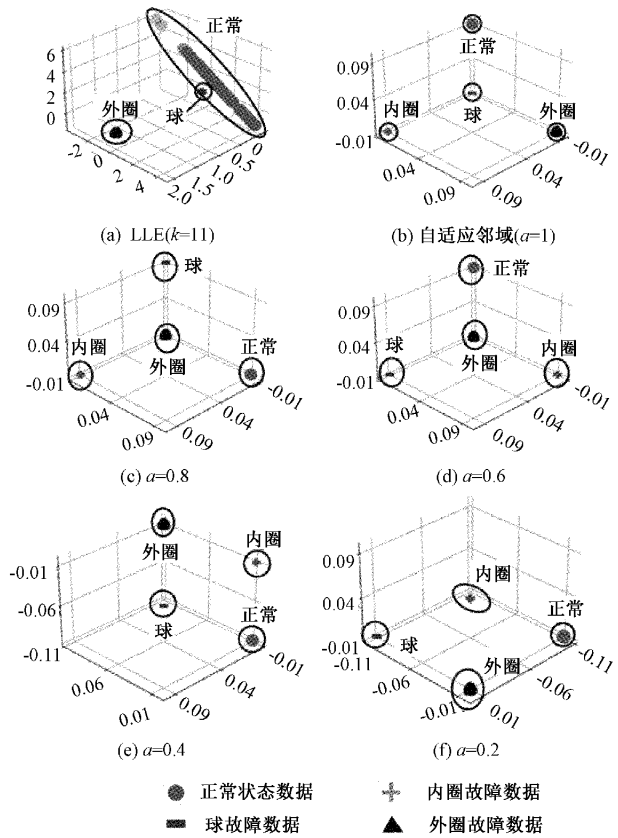

图 9 数据集 1 在不同权衡参数 α 下的可视化实验结果

由可视化实验结果图可知,不同的权衡参数 α 都对应优异的可视化结果,说明了 MIF-ALLE 算法对权衡参数 α 的选取不敏感,具有较好的鲁棒性。

其次,采用 Fisher 指标对实验结果进行定量聚类评价,进一步分析在权衡参数 $\alpha \in [0,1]$ 的情况下对算法的影响,此准则能够反映样本点之间的类内紧凑性以及类间

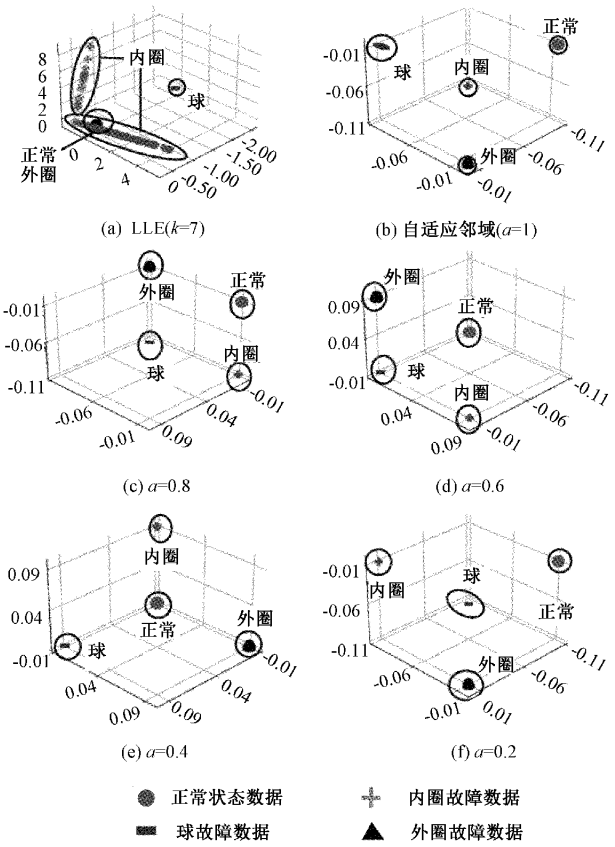

图 10 数据集 2 在不同权衡参数 α 下的可视化实验结果

分离度。Fisher 指标定义如下:

$$F = \frac{\mathrm{tr}(\boldsymbol{S}_b)}{\mathrm{tr}(\boldsymbol{S}_w)} \tag{16}$$

在式(16)中,tr(·)代表矩阵的迹,$\boldsymbol{S}_b$ 代表类间离散度,$\boldsymbol{S}_w$ 代表类内离散度,F 的值与聚类效果的好坏成正比。分别计算两个数据集在不同权衡参数 α 下所对应的 Fisher 指标,结果如图 11 所示。

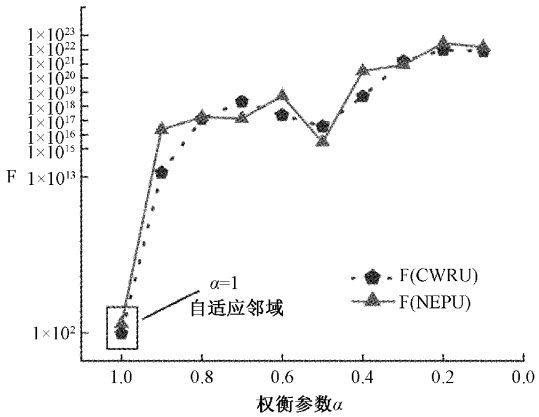

图 11 两个数据集在不同权衡参数 α 下对应的 Fisher 指标

分析图 11 可知,当 $\alpha = 1$ 时,只进行了自适应邻域的改进,虽然可视化降维效果较好,但是相较于其他 α 值(α = (0.9,0.8,…,0.1)),Fisher 指标大幅度降低,进一步说明

了多信息融合能够挖掘更全面的流形结构。

4　结　　论

本文提出了一种基于多信息融合的自适应局部线性嵌入算法。此算法利用切空间近似判据自适应选择邻域参数，解决了 LLE 算法邻域参数敏感性问题；将局部邻域中蕴含的局部切空间角度信息与局部线性信息相融合，解决了 LLE 算法挖掘结构单一问题。在两个轴承数据集上的实验表明，本文提出算法的具有优异的可视化效果、聚类性能以及鲁棒性。然而，此算法计算复杂度较高，只适用于处理固定工况下的数据，后续将对算法的轻量化以及可迁移性进行研究。

参考文献

[1] 陈仁祥，朱玉清，胡小林，等. 自适应正则化迁移学习的不同工况下滚动轴承故障诊断[J]. 仪器仪表学报，2021，41(8)：95-103.

[2] 李红月，高英杰，朱文昌. IAO 优化 SVM 的电机滚动轴承故障诊断[J]. 电子测量技术，2022，45(10)：126-132.

[3] 沈长青，雷飘，冯毅雄. 基于自适应流形嵌入动态分布对齐的轴承故障诊断[J]. 电子测量与仪器学报，2021，35(2)：33-40.

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

[5] 孙锐，王旭，张东东. 采用局部线性嵌入的稀疏目标跟踪方法[J]. 电子测量与仪器学报，2017，31(8)：1312-1320.

[6] CHANG H，YEUNG D Y. Robust locally linear embedding[J]. Pattern Recognition，2006，39(6)：1053-1065.

[7] HU W M，GAO J，WU O，et al. Anomaly detection using local kernel density estimation and context-based regression[J]. IEEE Transactions on Knowledge and Data Engineering，2020，32(2)：218-233.

[8] XU K P，CHEN L F，WANG S R. Data-driven kernel subspace clustering with local manifold preservation[J]. IEEE International Conference on Data Mining Workshops，2022：876-884.

[9] 杨静林，唐林波，宋丹. 基于自适应聚类流形学习的增量样本降维与识别[J]. 系统工程与电子技术，2015，37(1)：199-205.

[10] WANG B Y，HU Y L，GAO JB，et al. Learning adaptive neighborhood graph on grassmann manifolds for video/image-set subspace clustering[J]. IEEE Transactions on Multimedia，2021，DOI：10.1109/TMM.2020.2975394.

[11] LI X Y，FAN H，LIU J L，et al. One-step unsupervised clustering based on information theoretic metric and adaptive neighbor manifold regularization[J]. Engineering Applications of Artificial Intelligence，2023，DOI：10.1016/j.engappai.2023.105880.

[12] 梅松青，周洪建. 基于信息熵的局部线性嵌入[J]. 计算机工程与科学，2014，36(9)：1806-1811.

[13] SHENG Y P，WANG M，WU T X，HAN X. Adaptive local learning regularized nonnegative matrix factorization for data clustering[J]. Applied Intelligence，2019，49(6)：2151-2168.

[14] ZHAO P，WU H J，HUANG S D. Multi-view graph clustering by adaptive manifold learning[J]. Mathematics，2022，10(11)：1821.

[15] LIU Q，HE H，LIU Y，et al. Local linear embedding algorithm of mutual neighborhood based on multi-information fusion metric[J]. Measurement，2021，DOI：10.1016/j.measurement.2021.110239.

[16] ZHANG Y，ZHANG R，GAO Z W，et al. Dual-weight local linear embedding algorithm based on adaptive neighborhood[J]. Transactions of the Institute of Measurement and Control，2022，45(8)：1411-1421.

[17] MUHAMMAD Z H S，ZAHOOR A，HU L S. Weighted linear local tangent space alignment via geometrically inspired weighted pca for fault detection[J]. IEEE Transactions on Industrial Informatics，2023，19(1)：210-219.

[18] YIN J，SUN S. Incomplete multi-view clustering with cosine similarity[J]. Pattern Recognition，2022，DOI：10.1016/j.patcog.2021.108371.

作者简介

刘庆强，副教授，硕士，主要研究方向为信息安全、智能控制、信号处理与故障诊断。

E-mail：petroboy@163.com

魏朝阳（通信作者），硕士研究生，主要研究方向为轴承故障诊断。

E-mail：2662889018@qq.com