

利用测地线距离的改进谱聚类算法

杨清宇^{1,2}, 孙凤伟², 张墨², 张迪², 庄健³

(1. 西安交通大学机械制造系统工程国家重点实验室, 710049, 西安;

2. 西安交通大学电子与信息工程学院, 710049, 西安;

3. 西安交通大学机械工程学院, 710049, 西安)

摘要: 针对往复式压缩机故障数据空间分布复杂、常规算法不能有效聚类的问题,提出了一种改进的谱聚类算法. 该算法使用新的相似度矩阵计算方式,根据故障数据流形分布的特点引入测地线距离取代欧氏距离作为数据间的关系度量;通过计算各数据点的邻域密度因子有效地识别和剔除了噪声点;利用基于密度的局部欧氏距离调整方法对流形间隙过小的区域进行了处理. 在几个人工数据集和往复式压缩机故障数据集上的测试结果表明,改进谱聚类算法对于具有流形分布、多尺度、有噪声、流形间隙过小甚至交叉等特点的数据具有很好的聚类能力,聚类准确率比常规的 k -均值和 MSCA 谱聚类算法分别提高了 50.86% 和 8.6%.

关键词: 往复式压缩机;谱聚类;测地线距离;欧氏距离调整

中图分类号: TP18 **文献标志码:** A **文章编号:** 0253-987X(2012)08-0001-07

An Improved Spectral Clustering Algorithm Using Geodesic Distance

YANG Qingyu^{1,2}, SUN Fengwei², ZHANG Zhao², ZHANG Di², ZHUANG Jian³

(1. State Key Laboratory for Manufacturing Systems Engineering, Xi'an Jiaotong University, Xi'an 710049, China;

2. School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China;

3. School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: An improved spectral clustering algorithm is proposed to focus on the problem that the general clustering algorithms are invalid for reciprocating compressor fault data lying on complex manifold. A new affinity matrix is obtained. The geodesic distance replaces the traditional Euclidian distance to measure the similarity of data, and neighborhood-based density factor is used to identify and to remove noise points. Moreover, density-based local Euclidian distance adjustment is introduced into areas with small gap between manifolds. The proposed method is implemented on several artificial datasets and a real reciprocating compressor fault dataset. Experimental results show that the new algorithm can accomplish the clustering for data with noise and multi-scale character, especially when the manifolds have small gaps or crossover between each other. Its accuracy is 50.86% and 8.6% higher than those of k -means and MSCA respectively.

Keywords: reciprocating compressor; spectral clustering; geodesic distance; Euclidian distance adjustment

故障诊断通过分析数据内隐含的状态特征,使每一个数据样本对应一种设备工况,从而实现对设

备正常工况和故障工况的识别. 故障诊断实际上是要解决故障数据分类/聚类问题^[1],本质上属于模式

收稿日期: 2012-03-05. 作者简介: 杨清宇(1974—),男,副教授,博士生导师. 基金项目: 国家自然科学基金资助项目(61075001).

网络出版时间: 2012-05-11

网络出版地址: <http://www.cnki.net/kcms/detail/61.1069.T.20120511.1140.001.html>

识别问题^[2]. 聚类作为数据挖掘和模式识别的一种重要方法,在故障诊断领域得到了广泛的应用^[3-8].

常用的聚类方法有 k -均值聚类、Bayesian 聚类等,这些方法理论可靠、算法简洁、收敛速度快,在处理特征空间为球体或者超球体的数据时具有较好的聚类效果,但对于流形分布等复杂数据集往往效果不佳. 谱聚类算法以谱图理论为依据,本质上是将聚类转化为图的最优分割问题,此时聚类等同于基于一定分割准则把图划分为不同的子图的过程,每个子图代表一类. 谱聚类算法具有适用于分布不规则的非凸数据、无高维数据引起的奇异性、可收敛到全局最优等特点,较其他聚类算法更适用于压缩机数据. Ng 等人提出的 NJW (Ng, Jordan, Weiss) 算法^[9]是一种具有代表性的谱聚类算法,对复杂分布的数据具有较好的聚类效果. Guido 等人在此基础上提出了可以自动确定类数的算法 MSCA (modified spectral clustering algorithms)^[10],使用更加方便. 虽然 NJW 算法及其改进算法 MSCA 在处理复杂分布数据时有其独到的优势,但由于多级往复式压缩机故障数据往往分布在不规则的空间内,且不同故障类数据间有一定程度的重叠,因此将其用于往复式压缩机故障诊断时仍需进行针对性改进.

在处理复杂空间分布的数据集方面,出现了多种基于核函数的聚类算法^[11-12],这些算法通过非线性映射将低维线性不可分的原始数据投影到高维空间内,使得数据在高维空间内线性可分. 这些算法虽然性能优良,但使用的核函数多依据经验选择,没有统一的选择标准,需要针对具体的数据分布特点进行选择,核函数的选取对聚类结果影响较大.

为此,本文在 NJW 算法及其改进算法 MSCA 的基础上,针对往复式压缩机故障数据分布特点,提出了一种新的相似度矩阵计算方法,该方法直接在原始数据空间内进行计算. 通过几个有代表性的人工数据集的聚类结果以及一个真实的往复式压缩机故障诊断实验结果可以看出,本文算法与其他聚类算法相比有较高的准确率.

1 NJW 谱聚类算法和 MSCA 回顾

对于 N 个 d 维向量构成的数据集 $\mathbf{X} \in \mathbf{R}_{N \times d}$, NJW 谱聚类算法步骤如下^[9].

步骤 1 构造相似度矩阵 $\mathbf{A} = (a_{ij})_{N \times N}$, 其中 $a_{ij} = \exp\left(-\frac{d^2(\mathbf{x}_i, \mathbf{x}_j)}{\delta^2}\right)$; $d(\mathbf{x}_i, \mathbf{x}_j)$ 为点 $\mathbf{x}_i, \mathbf{x}_j$ 间的欧氏距离; δ 是尺度参数,控制相似度随距离增大而减小的

速度.

步骤 2 计算拉普拉斯矩阵 $\mathbf{L} = \mathbf{B}^{-1/2} \mathbf{A} \mathbf{B}^{-1/2}$, 其中 $\mathbf{B} = \text{diag}\left(\sum_{j=1}^d a_{ij}\right)$ 为对角度矩阵.

步骤 3 计算矩阵 $\mathbf{L}$ 的特征向量和特征值,将其特征值降序排列,提取前 q 个特征值所对应的特征向量,构造矩阵 $\mathbf{Y} \in \mathbf{R}_{N \times q}$. 将 $\mathbf{Y}$ 按行规范化,得到规范化的矩阵 $\mathbf{Y}' = (y'_{ij})_{N \times q}$, 其中 $y'_{ij} = \frac{y_{ij}}{\left(\sum_{j=1}^q y_{ij}^2\right)^{1/2}}$.

步骤 4 将 $\mathbf{Y}'$ 的每一行视为一个 q 维数据,利用 k -均值或者其他的聚类方法对 $\mathbf{Y}'$ 进行聚类.

步骤 5 若 $\mathbf{Y}'$ 中的点 $\mathbf{y}_i'$ 划分到第 j 类,则将 $\mathbf{x}_i$ 也对应地划分到第 j 类.

NJW 算法无法自动确定聚类数目,需要用户自己指定. 为此,Guido 等人提出了一种可以自动确定聚类数目的改进谱聚类算法——MSCA 算法^[10]. MSCA 算法使用如下步骤取代 NJW 算法的步骤 3 和步骤 4.

步骤 1 根据 NJW 算法首先计算矩阵 $\mathbf{L}$,并初始化 $q=2$.

步骤 2 取矩阵 $\mathbf{L}$ 的前 q 个特征向量构造矩阵 $\mathbf{Y} \in \mathbf{R}_{N \times q}$,将 $\mathbf{Y}$ 按行规范化,得到矩阵 $\mathbf{Y}' \in \mathbf{R}_{N \times q}$.

步骤 3 将 $\mathbf{Y}'$ 的每一行视为一个 q 维数据,在该 q 维的空间中初始化 $q+1$ 个中心点 $\mathbf{c}_i$, 其中 $i=1, 2, \dots, q+1$,第 $q+1$ 个中心点初始化为原点.

步骤 4 $\mathbf{Y}'$ 中每个点 $\mathbf{y}_j'$ 到每个中心点 $\mathbf{c}_i$ 的距离 $d_1(\mathbf{y}_j', \mathbf{c}_i)$ 由下式计算

$$d_1(\mathbf{y}_j', \mathbf{c}_i) = \begin{cases} (\mathbf{y}_j' - \mathbf{c}_i)^T \mathbf{M} (\mathbf{y}_j' - \mathbf{c}_i), & \mathbf{c}_i^T \mathbf{c}_i > \varepsilon \\ d(\mathbf{y}_j', \mathbf{c}_i), & \mathbf{c}_i^T \mathbf{c}_i \leq \varepsilon \end{cases} \quad (1)$$

式中: $d(\mathbf{y}_j', \mathbf{c}_i)$ 是 $\mathbf{y}_j'$ 与 $\mathbf{c}_i$ 之间的欧氏距离; $\mathbf{c}_i^T \mathbf{c}_i$ 为 $\mathbf{c}_i$ 与原点的距离; ε 为预先指定的参数; $\mathbf{M} = \frac{1}{\lambda} \left(\mathbf{I}_q - \frac{\mathbf{c}_i \mathbf{c}_i^T}{\mathbf{c}_i^T \mathbf{c}_i} \right) + \lambda \frac{\mathbf{c}_i \mathbf{c}_i^T}{\mathbf{c}_i^T \mathbf{c}_i}$, λ 为延伸系数,控制 $\mathbf{c}_i$ 邻域的拉伸程度.

步骤 5 在上述度量下将每个点 $\mathbf{y}_j'$ 划分到距其最近的中心点所在的类,计算每一类数据点的均值,并作为新的中心点.

步骤 6 返回步骤 3 继续计算,直到中心点几乎不变或达到最大迭代次数,停止迭代.

步骤 7 如果第 $q+1$ 个类中有数据,则可以确定有一个新类,设 $q=q+1$,返回步骤 2 继续执行;否则,输出类的数目和每类中的数据,完成 $\mathbf{Y}'$ 的聚类.

2 基于测地线距离的改进谱聚类算法

在一个多级往复式压缩机故障诊断实验平台上采集故障数据,实验对象为一台 W-0.6/12.5-S 型两级三缸往复式压缩机,主要由两套一级气缸组件、一套二级气缸组件、驱动电机、传动和冷却系统以及储气罐等部件组成.实验平台如图 1 所示.

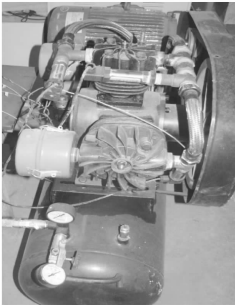


图 1 压缩机故障诊断实验平台

通过人为在阀片上打孔的方式模拟了正常、一级缸进气阀轻微泄漏、一级缸进气阀严重泄漏、二级缸排气阀轻微泄漏、二级缸排气阀严重泄漏 5 种工况,这 5 种工况分别用 F1~F5 来表示.以 200 Hz 的频率对 5 种工况采样 5 s,各采集 1 000 组数据,每组数据有 8 维,每一维对应一个参数,参数含义如表 1 所示.5 种工况的数据空间分布如图 2 所示.

表 1 参数含义表

维数	参数	参数含义
1	$T_1/^{\circ}\text{C}$	一级缸 A 排气温度
2	$T_2/^{\circ}\text{C}$	一级缸 B 排气温度
3	$T_3/^{\circ}\text{C}$	二级缸进气温度
4	$T_4/^{\circ}\text{C}$	二级缸排气温度
5	P_1/MPa	一级缸 A 排气压力
6	P_2/MPa	一级缸 B 排气压力
7	P_3/MPa	二级缸进气压力
8	P_4/MPa	二级缸排气压力

研究发现,往复式压缩机故障数据具有以下特点:①数据分布在非规则复杂空间内,并呈现一定的流形分布;②数据类间和类内具有多尺度特征,即不同类数据“紧密”程度不同,同类数据之间的相互距离也有变化;③由于数据本身的分布特点和数据集内噪声的影响,导致不同类数据所处的流形间隙过小,甚至有一定程度交叉.

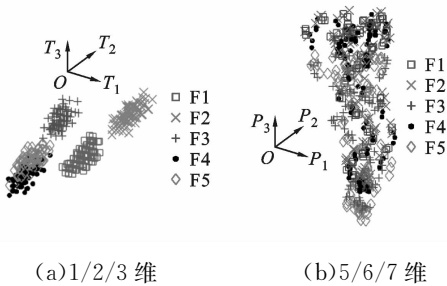


图 2 5 种工况下数据 1/2/3 维和 5/6/7 维空间分布

为此,本文选择了几个能较好地代表往复式压缩机故障数据特征的人工数据集,根据 MSCA 算法的聚类结果对其进行改进,并在往复式压缩机故障数据集上,对本文算法、MSCA 及 k -均值算法进行了对比分析.

2.1 基于测地线距离的故障数据相似性度量

往复式压缩机故障数据具有流形分布的特点,为此选择如图 3 所示的 4 个数据集,前 3 个数据集沿流形有不同程度的发散,呈现多尺度的特点,第 4 个数据集则同时具有流形分布、有噪声和多尺度的特点.为了比较流形距离与欧氏距离的不同,在图 3c 中选取了 1、2、3 三个点,其中 1、2 两点在不同流形上,属于不同的类,但相互之间距离较近;1、3 两点之间距离较远,但在同一流形上,属于相同的类.

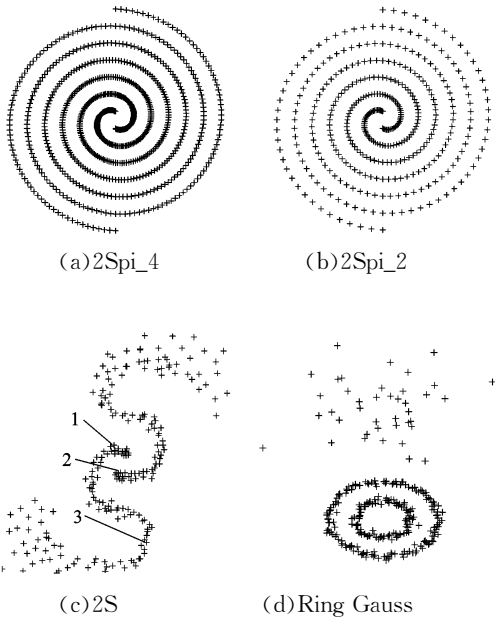


图 3 4 个人工数据集的原始数据分布

在常用的聚类算法中,大多采用欧氏距离对故障数据之间的关系进行度量.图 3c 中的 3 个点 1、2、3 若采用欧氏距离进行度量,则 1、2 距离最近,1、

3 距离较远. 为了能正确反映 3 个点的关系, 本文采用测地线距离进行故障数据的相似性度量, 取代 NJW 和 MSCA 算法中的欧氏距离度量. 测地线距离定义为^[6]

$$d_2(\mathbf{x}_i, \mathbf{x}_j) = \min_{p \in P_{ij}} \sum_{k=1}^n \sum_{l=1}^n u_{kl} d'(\mathbf{x}_k, \mathbf{x}_l) \quad (2)$$

$$\text{式中: } u_{kl} = \begin{cases} 1, & d'(\mathbf{x}_k, \mathbf{x}_l) \in p \\ 0, & \text{其他} \end{cases}, \text{其中 } d'(\mathbf{x}_k, \mathbf{x}_l) = \begin{cases} d(\mathbf{x}_k, \mathbf{x}_l), & \mathbf{x}_k \rightarrow \mathbf{x}_l, d(\mathbf{x}_k, \mathbf{x}_l) \text{ 为节点 } \mathbf{x}_k \text{ 与 } \mathbf{x}_l \\ +\infty, & \text{其他} \end{cases}$$

的欧氏距离, $\mathbf{x}_k \rightarrow \mathbf{x}_l$ 表示节点 $\mathbf{x}_l$ 是 $\mathbf{x}_k$ 的 K 个最近邻节点之一, P_{ij} 为连接点 $\mathbf{x}_i$ 与 $\mathbf{x}_j$ 的所有路径的集合, p 为一系列近邻节点间的边叠加生成的一条联通 $\mathbf{x}_i$ 与 $\mathbf{x}_j$ 的路径.

与欧氏距离相比, 测地线距离是由数据所在流形表面上的一系列短线段叠加而成, 这使得位于同一流形上的数据间的距离较小, 而不同流形上的数据间距离较大. Tenenbaum 等人在非线性降维的研究中, 给出了一种计算测地线距离的方法^[13].

使用欧氏距离的 MSCA 算法的聚类结果如图 4 所示.

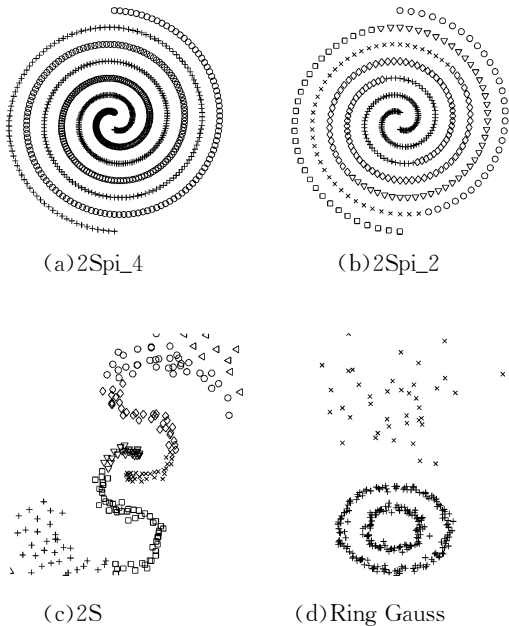


图4 使用欧氏距离的 MSCA 算法的聚类结果

由图 4 可知, 基于欧氏距离的 MSCA 算法能够正确聚类发散较小的 2Spi_4 数据集, 但对后 3 个数据集则无法正确聚类, 表明该方法对类内紧凑、类间有较大间隙的数据集有一定的聚类能力, 但对过于发散的、多尺度以及有噪声的数据集, 聚类能力

变差.

利用测地线距离的改进谱聚类算法的聚类结果如图 5 所示.

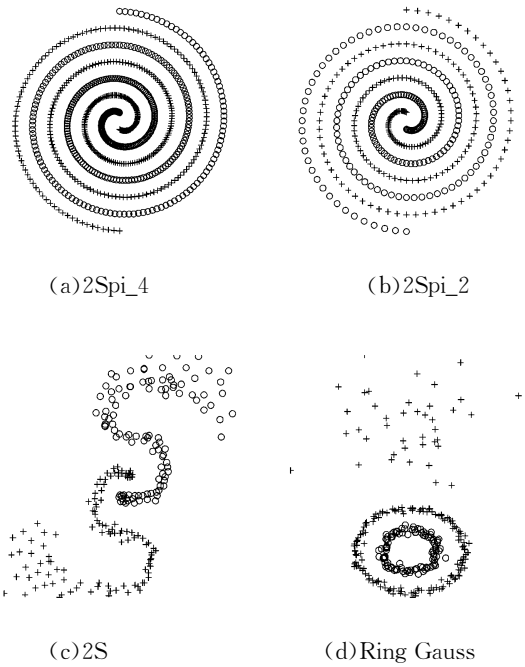


图5 利用测地线距离的改进算法的聚类结果

由图 5 可知, 基于测地线距离的谱聚类算法能够正确聚类前 3 个呈流形分布且具有多尺度变化的数据集, 但当数据集中存在较大噪声时, 该算法容易出现过分割和欠分割的情况. 因此, 需要针对噪声存在的情况对该算法进行改进.

2.2 噪声点的识别与剔除

由于往复式压缩机数据集中某些噪声点处于流形的间隙, 导致原本有较大间隙的 2 个流形很容易通过噪声点连接起来, 使得应用测地线距离度量后处于不同流形上的数据也会有较小距离, 导致 2 个流形被聚成同一类. 另有一些噪声点孤立存在于远离各个流形的位置, 在聚类时容易被划分为单独的类. 此处称这两类噪声点为“野点”. 为了正确聚类, 需要在计算测地线距离之前识别并剔除“野点”.

Zhou 等人^[14]在基于邻域的聚类算法中, 定义了邻域密度因子 (neighborhood-based density factor, NDF), 用以识别稠密点、平坦点和稀疏点, 这里将邻域密度因子应用于野点的识别和剔除. 邻域密度因子定义为

$$F_{ND}(k, \mathbf{x}) = \frac{|N_{Rk}(\mathbf{x})|}{|N_k(\mathbf{x})|} \quad (3)$$

式中: $F_{ND}(k, \mathbf{x})$ 为邻域大小为 k 时点 $\mathbf{x}$ 的邻域密度因子; $N_k(\mathbf{x})$ 和 $N_{Rk}(\mathbf{x})$ 分别为 $\mathbf{x}$ 的 k 邻域和反 k 邻域. 具体地, 记 $S_{kNN}(\mathbf{x})$ 为 $\mathbf{x}$ 的 k 近邻集合, 并令 $r = \max_{\mathbf{x} \in X} d(\mathbf{x}, S_{kNN}(\mathbf{x}))$, 则有

$$N_k(\mathbf{x}) = \{\mathbf{y} \in \mathbf{X} \mid d(\mathbf{y}, \mathbf{x}) \leq r, \mathbf{y} \neq \mathbf{x}\} \quad (4)$$

$$N_{Rk}(\mathbf{x}) = \{\mathbf{y} \in \mathbf{X} \mid \mathbf{x} \in N_k(\mathbf{y})\} \quad (5)$$

将 $F_{ND}(k, \mathbf{x}_i) < \mu$ 的点 $\mathbf{x}_i$ 标记为野点, 先将其从原始数据集中剔除, 再对剔除野点后的数据集进行聚类, 为了实现对所有数据的聚类, 聚类完成后可将野点划分到与其欧氏距离最近点所在的类内.

图6显示了基于测地线距离的改进谱聚类算法在剔除野点前后的聚类结果对比.

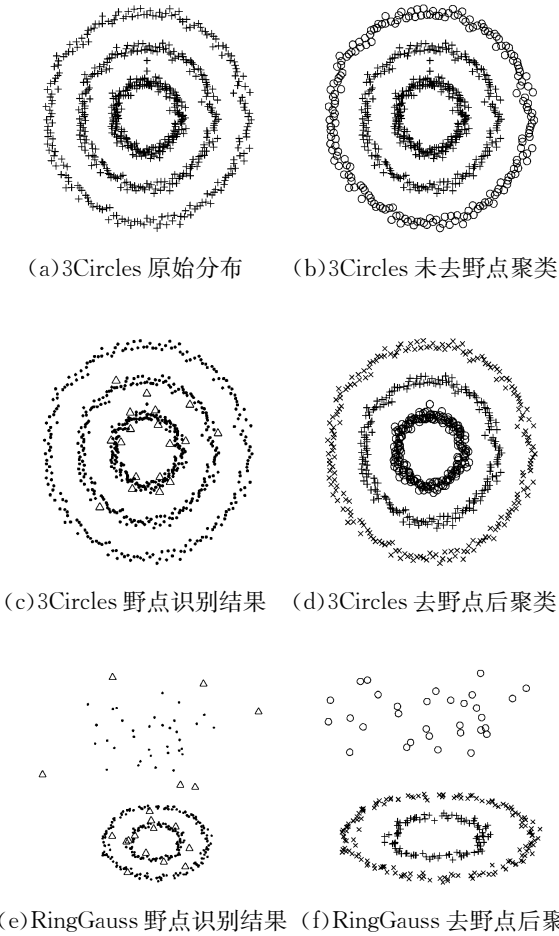


图6 改进算法在剔除野点前后的聚类结果对比

从图6可以看出, 使用NDF可有效识别出数据集内的噪声点, 在剔除野点后, 间隙内存在噪声的流形之间不易发生粘连, 同时空间内远离所有点分布的孤立点也不会再被聚成单独的类.

2.3 基于密度的局部欧氏距离调整

除了噪声外, 往复式压缩机故障数据在参数空

间内不同流形间会出现间隙过小甚至一定的交叉, 使用测地线距离度量会使这样的2个流形上的数据也具有较小的距离, 从而导致这2个流形被聚成同一类. 此外, 压缩机故障数据不同流形间有密度差异, 但同一流形上互为邻域的点密度差异较小, 为此本文利用了基于密度的局部欧氏距离调整方法, 调整后点 $\mathbf{x}_i, \mathbf{x}_j$ 之间的距离为

$$d_3(\mathbf{x}_i, \mathbf{x}_j) = a(\mathbf{x}_i, \mathbf{x}_j) d(\mathbf{x}_i, \mathbf{x}_j) \quad (6)$$

式中

$$a(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} 1 - \exp\left\{-\left(\frac{|\rho_i - \rho_j|}{\max_{1 \leq m \leq N} \rho_m}\right)^2\right\}, & \mathbf{x}_i \rightarrow \mathbf{x}_j \\ 1, & \text{其他} \end{cases} \quad (7)$$

其中 $\rho_i = \sum_{j=1}^N d(\mathbf{x}_i, \mathbf{x}_j)$ 反映了点 $\mathbf{x}_i$ 所处区域的密度, 此处也可选 k 邻域最大距离、 k 邻域平均距离等作为点 $\mathbf{x}_i$ 所处区域的密度.

由于 k 邻域不具有对称性, 因此上面得到的 $d_3(\mathbf{x}_i, \mathbf{x}_j)$ 也不具备对称性, 为了便于后续谱聚类算法使用, 求得 $d_3(\mathbf{x}_i, \mathbf{x}_j)$ 的对称化形式

$$d'_3(\mathbf{x}_i, \mathbf{x}_j) = \frac{d_3(\mathbf{x}_i, \mathbf{x}_j) + d_3(\mathbf{x}_j, \mathbf{x}_i)}{2} \quad (8)$$

图7显示了在使用基于密度的局部欧氏距离调整前和调整后的聚类结果对比.

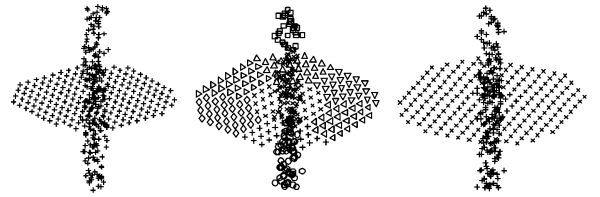


图7 局部欧氏距离调整前后的聚类结果

由图7可知, 经过局部欧氏距离调整后, 在处理流形交叉的情况能够达到更好的聚类效果, 由于2个流形的密度不一致, 通过基于密度的局部欧氏距离调整, 使交叉区域内点与点之间的欧氏距离产生不同程度的伸缩, 即对于每个点及其邻域内的点, 两者密度越相近则它们之间的欧氏距离缩小程度越大, 每个点与其邻域外的点之间的欧氏距离不作调整. 在交叉区域内, 算法减小了同类数据间的距离, 不同类数据间的距离保持不变, 等效于拉大了不同流形间隙, 对于流形交叉的情况起到了一定的作用.

2.4 改进谱聚类算法

本文针对往复式压缩机故障数据分布特点,引入测地线距离、噪声点识别剔除、基于密度的局部欧氏距离调整,相似度矩阵 $\mathbf{A}$ 计算流程如下.

步骤 1 计算欧氏距离矩阵 $\mathbf{D}$, 其中 $d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j)$ 为点 $\mathbf{x}_i, \mathbf{x}_j$ 间的欧氏距离.

步骤 2 根据式(3)计算每个点的邻域密度因子, 将 $F_{ND}(k, \mathbf{x}_i) < \mu$ 的点 $\mathbf{x}_i$ 标记为野点, 并将其从原始数据集 $\mathbf{X}$ 和矩阵 $\mathbf{D}$ 中剔除.

步骤 3 使用基于密度的局部欧氏距离调整算法, 由式(8)计算局部调整欧氏距离矩阵 $\mathbf{D}_3'$.

步骤 4 根据式(2)计算测地线矩阵 $\mathbf{D}_2$.

步骤 5 计算相似度矩阵 $\mathbf{A}$, $\mathbf{A}$ 中元素 $a_{ij} = 1/(d(\mathbf{x}_i, \mathbf{x}_j) + \epsilon)$, ϵ 为一个很小的正实数, 防止分母为零.

由此可知: 欧氏距离矩阵计算和基于密度的局部欧氏距离调整都通过两层循环计算一个 $n \times n$ 维矩阵, 计算复杂度均为 $O(n^2)$; 计算 NDF 的复杂度为 $O(mn)^{[14]}$; 计算测地线距离矩阵时, 若使用 Floyd 算法求取最短路径, 则其计算的复杂度为 $O(n^3)$, 若使用 Fibonacci heap 实现优先队列的 Dijkstra 算法, 则复杂度为 $O(nlgn + e)$, 其中 e 为邻接图的边数. 改进算法的计算复杂度主要由测地线距离矩阵的计算量决定, 至多为 $O(n^3)$, 而 NJW、MSCA 算法的复杂度主要由特征分解决定, 复杂度为 $O(n^3)^{[15]}$, 因此本文算法与 NJW、MSCA 算法相比, 计算复杂度在同一数量级上.

3 往复式压缩机故障数据聚类实验

在 CPU 为 E6500 (双核 2.93 GHz)、内存为 2GB 的计算机上, 分别用本文算法、MSCA 和 k -均值算法对采集到的往复式压缩机故障数据进行聚类实验, 各算法的运行时间和数据聚类正确率如表 2 所示.

表 2 各算法的运行时间和数据聚类正确率对比

算法	运行时间/s	聚类正确率/%				
		F1	F2	F3	F4	F5
本文算法	5.25	87.3	92.5	100	98.7	100
MSCA	3.54	81.2	75.5	100	83.7	95.1
k -均值	0.15	19.5	49.9	85.2	40.5	29.1

由表 2 可知, 本文算法聚类效果最好, MSCA 次之, k -均值算法最差. 本文算法的聚类正确率相比

k -均值算法和 MSCA 分别提高了 50.86% 和 8.6% (5 种工况提高程度的平均值). 运行时间方面, 本文算法虽消耗时间最多, 但仍与 MSCA 在同一数量级上. 由于一级缸进气阀严重泄漏故障与其他工况特征差异明显, 其数据在空间内远离其他类, 所以 3 种算法都有较好的聚类效果. 对于其余 4 种工况, 本文算法和 MSCA 聚类效果明显好于 k -均值算法, 这是由于本文算法和 MSCA 对于流形分布的数据都具有一定的聚类能力, 而基于欧氏距离度量的 k -均值算法只能处理球形分布的数据, 对于呈现流形分布特征的数据无法正确聚类. 此外, 正常工况和两种轻微泄露故障之间特征差异较小, 三者的流形存在间隙过小甚至交叉的情况, 使得 MSCA 对于不同流形交界处的数据无法准确区分.

本文在谱聚类算法中引入测地线距离后, 针对流形间隙过小和交叉的问题使用了基于密度的局部欧氏距离调整, 等效“拉伸”了交界区域不同流形的间隙, 同时识别并剔除了噪声点, 使本文算法对故障数据的聚类效果明显好于 MSCA 和 k -均值算法.

4 结 论

本文在谱聚类算法的基础上, 提出了一种新的相似度矩阵计算方法, 先引入测地线距离作为数据间的关系度量, 又增加了野点识别与剔除、基于密度的局部欧氏距离调整. 测试结果表明, 本文算法对于具有流形分布、多尺度、有噪声、流形间隙过小甚至交叉的数据具有很好的聚类能力, 但由于算法中涉及图的最短路径求取、矩阵求逆、求特征值与特征向量等运算, 当数据集过大时, 易出现系统资源占用高、运算时间长的问題, 有待进一步改进.

参考文献:

[1] CHIANG L H, RUSSELL E L, BRAATZ R D. Fault detection and diagnosis in industrial systems [M]. London, Great Britain: Springer-Verlag London Limited, 2001.

[2] 屈梁生, 张海军. 机械诊断中的几个基本问题[J]. 中国机械工程, 2000, 11(2): 211-216.

QU Liangsheng, ZHANG Haijun. Some basic problems in machinery diagnostics[J]. China Mechanical Engineering, 2000, 11(2): 211-216.

[3] 王志华. 基于模式识别的柴油机故障诊断技术研究[D]. 武汉: 武汉理工大学, 2004.

[4] SUI Wentao, LU Changhou, ZHANG Dan. Bearing fault diagnosis based on feature weighted FCM cluster

- analysis[C] // Proc of ICCSSE. Piscataway, NJ, USA; IEEE, 2008; 518-521.
- [5] 张育林, 庄健, 王娜, 等. 一种自适应局部线性嵌入与谱聚类融合的故障诊断方法[J]. 西安交通大学学报, 2010, 44(1): 77-82.
ZHANG Yulin, ZHUANG Jian, WANG Na, et al. Fusion of adaptive local linear embedding and spectral clustering algorithm with application to fault diagnosis [J]. Journal of Xi'an Jiaotong University, 2010, 44(1): 77-82.
- [6] LI Gang, ZHUANG Jian, HOU Hongning, et al. A genetic algorithm based clustering using geodesic distance measure[C] // Proc of ICICIS. Piscataway, NJ, USA; IEEE, 2009; 274-278.
- [7] BAHRAMPOUR S, MOSHIRI B, SALAHSHOOR K. Weighted and constrained possibilistic C-means clustering for online fault detection and isolation[J]. Applied Intelligence, 2011, 35(2): 269-284.
- [8] DING Shifei, ZHANG Liwen, ZHANG Yu. Research on spectral clustering algorithms and prospects [C] // Proc of ICCET. Piscataway, NJ, USA; IEEE, 2010; 149-153.
- [9] NG A Y, JORDAN M I, WEISS Y. On spectral clustering: analysis and an algorithm [C] // Proc of ANIPS. Cambridge, MA, USA; MIT Press, 2002; 849-856.
- [10] SANGUINETTI G, LAIDLER J, LAWRENCE N. Automatic determination of the number of clusters using spectral algorithms[C] // Proc of IEEE Workshop on MLSP. Piscataway, NJ, USA; IEEE, 2005; 28-30.
- [11] 张莉, 周伟达, 焦李成. 核聚类算法[J]. 计算机学报, 2002, 25(6): 587-590.
ZHANG Li, ZHOU Weida, JIAO Licheng. Kernel clustering algorithm[J]. Chinese Journal of Computers, 2002, 25(6): 587-590.
- [12] DANIEL G, WITOLD P. Kernel-based fuzzy clustering and fuzzy clustering: a comparative experimental study [J]. Fuzzy Sets and Systems, 2010, 161(4): 522-543.
- [13] TENENGAUM J B, SILVA V D, LANGFORD J C. A global geometric framework for nonlinear dimensionality reduction[J]. Science, 2000, 290(5500): 2319-2323.
- [14] ZHOU Shuigeng, ZHAO Yue, GUAN Jihong, et al. A neighborhood-based clustering algorithm[J]. Lecture Notes in Computer Science, 2005, 3518(1): 361-371.
- [15] CHEN Wen-Yen, SONG Yangqiu, BAI Hongjie, et al. Parallel spectral clustering in distributed systems [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(3): 568-586.

[本刊相关文献链接]

- 张慧, 韩崇昭, 闫小喜. 概率假设密度滤波的谱聚类目标状态提取方法. 2012, 46(2): 1-5.
- 王翔, 郑建国. 求解约束优化问题的多成员人工蜂群算法. 2012, 46(2): 38-44.
- 刘弹, 杨景明, 罗爱玲. 具有全局聚类的多属性离散化算法. 2011, 45(9): 1-5.
- 王羨慧, 覃征, 张选平, 等. 采用仿射传播的聚类集成算法. 2011, 45(8): 1-5.
- 王翔, 郑建国, 张超群, 等. 采用约束蚁群优化的贝叶斯网络结构学习算法. 2011, 45(8): 54-61.
- 彭柳青, 张军英. 一种鲁棒的子空间聚类算法. 2011, 45(6): 13-19.
- 丁晓剑, 赵银亮. 优化极限学习机的序列最小优化方. 2011, 45(6): 7-12.
- 高羨明, 洪军, 刘瑞萍, 等. 多准则数控机床模块谱系聚类的划分方法研究. 2011, 45(5): 131-136.
- 马超, 陈西宏, 徐宇亮, 等. 广义邻域粗集下的集成特征选择及其选择性集成算法. 2011, 45(6): 34-39.
- 国强, 王常虹, 李峥. 支持向量聚类联合类型熵识别的雷达信号分选方法. 2010, 44(8): 63-67.
- 贾建华, 焦李成, 柳炳祥. 图像分割的谱聚类集成算法. 2010, 44(6): 93-98.
- 张兆军, 冯祖仁, 任志刚. 采用序优化的改进蚁群算法. 2010, 44(2): 15-19.
- 张育林, 庄健, 王娜, 等. 一种自适应局部线性嵌入与谱聚类融合的故障诊断方法. 2010, 44(1): 77-82.
- 王娜, 杜海峰, 王孙安. 一种基于流形距离的迭代优化聚类算法. 2009, 43(5): 76-79.
- 张云, 冯博琴. 利用标签的层次化搜索结果聚类方法. 2009, 43(4): 18-21.
- 李静耘, 杜正春, 楚国莉, 等. 基于聚类的多运行方式下电力系统稳定器设计. 2008, 42(2): 204-208.
- 张选平, 祝兴昌, 马琮. 一种基于边界识别的聚类算法. 2007, 41(12): 1387-1390.
- 张云, 冯博琴, 麻首强, 等. 蚁群-遗传融合的文本聚类算法. 2007, 41(10): 1146-1150.

(编辑 刘杨)