

一种自适应局部线性嵌入与谱聚类 融合的故障诊断方法

张育林, 庄健, 王娜, 王孙安
(西安交通大学机械工程学院, 710049, 西安)

摘要: 针对数据维数高、非线性且从高维观测空间分析数据模式困难的问题, 将改进的流形学习算法引入到数据聚类中, 提出了一种结合自适应局部线性嵌入和递归调用规范切融合的新方法. 采用自适应局部线性嵌入对原始数据进行非线性降维, 应用递归调用规范切对低维空间数据进行聚类, 通过对3组UCI标准测试数据集的仿真实验表明, 新方法能够将高维数据有效地映射到低维本质空间, 克服了传统方法对数据集结构的依赖性, 从而显著提高了谱聚类算法分类的准确性和稳定性. 同时, 对于田纳西-伊斯曼过程的数据实验, 表明了该方法对故障模式识别的可行性和有效性.

关键词: 局部线性嵌入; 谱聚类; 递归调用规范切; 故障诊断

中图分类号: TP183 **文献标志码:** A **文章编号:** 0253-987X(2010)01-0077-06

Fusion of Adaptive Local Linear Embedding and Spectral Clustering Algorithm with Application to Fault Diagnosis

ZHANG Yulin, ZHUANG Jian, WANG Na, WANG Sun'an
(School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: Focusing on the data with high dimensions and nonlinearity, in pattern recognition in high dimensional observation space, an improved manifold learning algorithm is introduced, and a new approach is proposed by combining adaptive local linear embedding (ALLE) and recursively applying normalized cut algorithm (RANCA). The adaptive local linear embedding algorithm is employed for nonlinear dimension reduction of original dataset, then recursively applying normalized cut algorithm is used in clustering of low dimensional data. The simulation results of three UCI standard datasets show that the new method can map high-dimensional data into low-dimensional intrinsic space successfully, solves the more dependence on the structure of datasets in the traditional methods, and the classification accuracy and robustness of spectral clustering algorithm are remarkably improved. The experiment results on tennessee-eastman process (TEP) also demonstrate the feasibility and effectiveness of the new method in fault pattern recognition.

Keywords: local linear embedding; spectral clustering; recursively applying normalized cut; fault diagnosis

随着系统的复杂化, 现代故障诊断采用解析模型法越来越困难, 因此基于信号处理和基于知识的方法成为目前研究的主要方向^[1]. 故障诊断本质上是典型的模式识别问题^[2], 即根据数据的内在特征,

对正常和各种故障状态进行分类. 聚类分析是模式识别等领域的一个重要方法, 其主要思想是将数据分类, 使得同类数据具有较高的相似度, 而类间数据的相似度则较小. 传统的聚类方法, 如 k-means、EM

(Expectation Maximization)等算法仅在具有凸形结构的数据上有好的效果,为了解决这个问题,学者们提出了谱聚类算法,在谱图理论的基础上,将聚类问题转化为图的最优划分问题^[3].

在故障诊断中,高维变量和大量相似监测数据增加了本质特征提取的难度,Seung^[4]等人认为感知可能以流形方式存在,对于高维观测数据集对象,其数据变量可以用少量几个影响因素来表示.降维可以使故障诊断模型结构变得简洁,从而提高了故障诊断的正确率和效率.传统的主成分分析(PCA)方法对线性数据有较好的处理效果,但对非线性结构数据的处理效果却欠佳. Roweis^[5]提出了局部线性嵌入(LLE)算法,使得流形学习迅速发展,LLE算法是把高维、非线性数据映射到全局的低维空间中,因此能够有效地发现非线性高维数据的某种本质变化规律.

本文提出一种基于自适应局部线性嵌入和谱聚类融合的方法,首先借鉴LLE算法,将高维数据进行非线性特征提取,然后针对LLE算法对不理想数据较为敏感的问题,提出了自适应局部线性嵌入算法.该算法借助于谱聚类算法中的规范切方法,采用递归调用规范切算法对数据进行聚类实验,解决了常规谱聚类无法从局部到整体分析数据内在规律的问题.

1 局部线性嵌入算法

LLE算法是一种非线性降维的算法,它采用局部线性逼近的思想,保持高维数据的局部几何关系,通过相互重叠的局部邻域数据,获得整体的结构.假设输入数据为采样于流形空间的 n 个 D 维实向量样本点 $\mathbf{x}_i, \mathbf{x}_i \in \mathbf{R}^D, i=1, 2, \dots, n$,输出为降维后的 n 个 d 维向量 $\mathbf{y}_i, \mathbf{y}_i \in \mathbf{R}^d, i=1, 2, \dots, n (d \ll D)$.算法的主要步骤如下.

步骤1 对于高维空间中的每个 $\mathbf{x}_i$,计算它和其余 $(n-1)$ 个样本点之间的欧氏距离,根据距离的远近,将最近的 k 个点作为邻域数据.

步骤2 用每个 $\mathbf{x}_i$ 的最近邻域线性组合近似表示 $\mathbf{x}_i$,通过最小化重构误差,得到局部重构权值矩阵

$$\boldsymbol{\varepsilon}(\mathbf{x}_i) = \sum_{j=1}^n \|\mathbf{x}_i - \sum_{j=1}^n \omega_{ij} \mathbf{x}_{ij}\|^2 \quad (1)$$

$$\sum_{j=1}^n \omega_{ij} = 1; \text{对于非 } \mathbf{x}_i \text{ 邻域的点, } \omega_{ij} = 0$$

式中: ω_{ij} 表示局部重构权值.

步骤3 根据 $\mathbf{x}_i$ 与邻域点 $\mathbf{x}_{ij}$ 的权值,通过求解稀疏

对称矩阵求出低维嵌入向量 $\mathbf{y}_i$,即

$$E(\mathbf{y}_i) = \sum_{i=1}^n \left\| \mathbf{y}_i - \sum_{j=1}^n \omega_{ij} \mathbf{y}_{ij} \right\|^2 \quad (2)$$

$$\sum_{i=1}^n \mathbf{y}_i = 0; \frac{1}{n} \sum_{i=1}^n \mathbf{y}_i \mathbf{y}_i^T = \mathbf{I}$$

式中: $E(\mathbf{y}_i)$ 为代价函数; $\mathbf{I}$ 为 $d \times d$ 单位矩阵.

2 自适应局部线性嵌入算法

由于LLE算法只有邻域数据参与重构,因此依赖于邻域的选取,要求对于流形空间的采样数据多,并且均匀或没有噪声.在LLE算法中,对 k 的选取非常重要,如果 k 选取得太小,就很难保证数据的整体几何性质,如果 k 选取得太大,可能会将流形空间相距较远的点化为邻域,从而扭曲了降维结果.因此,对于采样数据不均匀、稀疏或者含有噪声点的情况,邻域个数的选取尤为重要^[6].

图1为Swiss-roll理想采样和随机采样的3维数据,以及各自采用LLE算法降至2维的结果.可以看出,对于同样的采样数量和邻域个数,当采样数据理想时,LLE算法将源数据邻域映射到目标空间邻域,当采样数据不理想时,嵌入结果明显扭曲,因此LLE算法对于不理想数据的敏感性较强.

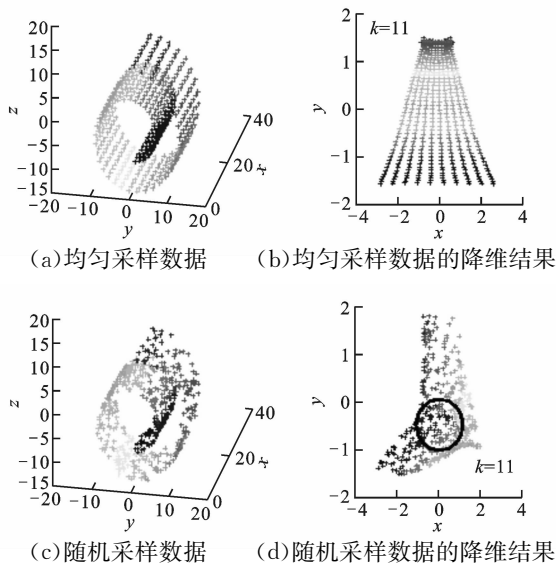


图1 Swiss-roll数据的降维结果

LLE算法选取的 k 不能小于高维数据的本质维数, k 一般取比 d 大一定值的数值,高维数据点之间要有一定的邻域重叠性,这样才能将 n 个子区域相关联,从而恢复数据的全局结构.但是,如果 k 取得太大时,也会导致错误的结果.当 k 太大时,会产生非唯一的重构权值,并且会将高维数据中非邻近点纳入到邻域中,从而无法保证局部线性结构,产生

错误的映射结果. 特别是对于不理想的采样数据, LLE 算法对于 k 的选取更加敏感, 本文根据流形数据各个区域的分布特点, 提出了自适应局部线性嵌入算法.

定义 1 设 $\mathbf{x}_{ij}$ 为 $\mathbf{x}_i$ 的向量邻域 $\mathbf{N}_i = \{\mathbf{N}(\mathbf{x}_i) | \mathbf{x}_i \in \mathbf{X}\}$ 中的第 j 个点, $j=1, 2, \dots, k$, 则 $\mathbf{x}_{ij}$ 的局部集群系数为

$$C_{ij} = \frac{2E_{ij}}{k(k-1)} \quad (3)$$

式中: E_{ij} 为 $\mathbf{x}_{ij}$ 与邻域点互为邻域的次数, 如果 $\mathbf{x}_{ij}$ 同时也是 $\mathbf{N}(\mathbf{x}_i)$ 的邻域, 则 $E_{ij} = E_{ij} + 1$.

定义 2 设 $\mathbf{x}_i (i=1, 2, \dots, n)$ 的初始 k 个邻域数据为 $\mathbf{x}_{ij}, j=1, 2, \dots, k, \mathbf{x}_i$ 与邻域数据的平均欧氏距离为

$$d_o = \sum_{j=1}^k \|\mathbf{x}_i - \mathbf{x}_{ij}\| / k, i=1, 2, \dots, n \quad (4)$$

流形邻域的平均距离为

$$d_M = \sum_{i=1}^n d_o / n \quad (5)$$

定义 3 自适应邻域个数为

$$k_i = k d_M / d_o \quad (6)$$

从定义 2 和定义 3 中可以看出, 当 $\mathbf{x}_i$ 邻域数据比较稀疏时, 平均邻域距离较大, 算法会自动降低 k 值, 从而避免了非近邻或者噪声作为重构数据造成的错误映射. 当 $\mathbf{x}_i$ 邻域数据比较致密时, 算法将自动增加 k 值, 避免了邻域间因数据缺乏流形关联, 而引起的整体结构扭曲. 从式(6)中可以看出, 自适应邻域个数通过分析采样数据的关联情况而设定邻域的大小, 为了满足文献[6]中的条件, 设定最大邻域为 $k_{\max}$ 和最小邻域为 $k_{\min}$.

将自适应局部线性嵌入算法描述如下: ①由 N 个 D 维实向量 $\mathbf{x}_i$ 构成高维数据向量集 $\mathbf{X} = \{\mathbf{x}_i | \mathbf{x}_i \in \mathbf{R}^D\}, i=1, 2, \dots, n$; ②嵌入空间维数 $d, k, k_{\max}, k_{\min}$ 和小世界选择概率 p ; ③ $\mathbf{y}_i, \mathbf{y}_i \in \mathbf{R}^d, i=1, 2, \dots, n$, 由 n 个 $\mathbf{y}_i$ 构成的数据集记作 $\mathbf{Y}$, 其中 $d \ll D$.

步骤 1 为每个样本点 $\mathbf{x}_i$ 寻找 k 个最近邻域数据 $\mathbf{x}_{ij}$ (数据集记作 $\mathbf{N}_i, j=1, 2, \dots, k$) 和 $k_{\max}$ 个邻域构成的 $\mathbf{N}_{\max}$, 其中 $k < k_{\max}$.

步骤 2 根据定义 2 和定义 3 计算 $\mathbf{x}_i$ 的邻域个数, 并设定 $k_{\min} < k_i < k_{\max}$.

步骤 3 根据式(1)计算 C_{ij} , 当 $k_i < k$ 时, 在 $\mathbf{N}_i$ 中剔除局部集群系数较小的 $p(k-k_i)$ 个点, 按照欧氏距离剔除其余 $(1-p)(k-k_i)$ 个距离较大的数据点. 当 $k_i > k$ 时, 保留 $\mathbf{N}_i$ 作为邻域数据, 并在 $(\mathbf{N}_{\max} -$

$\mathbf{N}_i)$ 中选择局部集群系数较大的 $p(k_i-k)$ 个点, 按照欧氏距离选择其余 $(1-p)(k-k_i)$ 个距离较小的数据点. 优化后的各点的邻域数据集记作 $\mathbf{N}_o(\mathbf{x}_i), i=1, 2, \dots, n$.

步骤 4 按照标准局部线性嵌入算法的步骤 2 和步骤 3, 来完成局部重构权值矩阵的计算和低维嵌入向量的计算, 以及降维过程.

3 ALLE 与谱聚类融合的聚类算法

3.1 递归调用规范切

递归调用规范切是一种经典的谱聚类算法. 在图分割聚类方法中, 借鉴层次凝聚的思想, 将图分割聚类问题转化为图的两划分问题, 即将图的顶点划分成为 2 个分离的子集 C, C' , 使划分具有最大组内相似性和最小组间相似性. Shi 等^[7] 提出的规范切判据为

$$\min N_c(C, C') = \frac{c_u(C, C')}{c_u(C, V)} + \frac{c_u(C, C')}{c_u(C', V)} \quad (7)$$

其中

$$c_u(C, C') = \sum_{i \in C, j \in C'} a_{ij}; c_u(C, V) = \sum_{i \in C, j \in V} a_{ij}; c_u(C', V) = \sum_{i \in C', j \in V} a_{ij}$$

定义 4 相似性矩阵 $\mathbf{A}(i, j) = a_{ij}$, 采用文献[3]中的定义, 相似度矩阵 $\mathbf{D}_u(i, j) = \sum_j a_{ij}$. Laplacian 矩阵 $\mathbf{L} = \mathbf{D}_u - \mathbf{A}$. 这里, 可以把图的两划分问题理解为求指示列向量 $\mathbf{y}, \mathbf{y}_i = 1, i \in C; \mathbf{y}_i = -1, i \in C'$. 根据 Rayleigh 定理^[8] 可知, 对应于 $\mathbf{D}_u^{-1}\mathbf{A}$ 的第 2 最大特征值的特征向量为式(7)最优解时, 即可有效找到划分数据.

规范切判据两划分数据的步骤可描述为: ① 计算 $\mathbf{A}, \mathbf{D}_u$; ② 计算 $\mathbf{P} = \mathbf{D}_u^{-1}\mathbf{A}$; ③ 对 $\mathbf{P}$ 进行特征分解, 令 $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n$ 为对应于 $1 = \lambda'_1 \geq \lambda'_2 \geq \dots \geq \lambda'_n$ 的特征向量, 求出第 2 最大特征值 λ'_2 对应的特征向量 $\mathbf{r}_2$; ④ 把 $\mathbf{r}_2$ 中的第 i 个元素和第 i 个数据点对应, 将 $\mathbf{r}_2$ 中各元素按递减顺序排列, 数据点也做相应的调整, 在 $i=1, 2, \dots, n-1$ 中寻找最优分割点 ($i=i_0$), 使得 $N_c(C_i, C'_i)$ 最小.

一般随着数据被不断划分, 相应的规范切判据值不断增大. 因此, 定义递归调用规范切算法停止划分的条件为: 每一类数据划分后所得到的规范切判据值大于或等于某一给定的阈值 threshold, 否则不断将两划分后的数据继续按照步骤 1~步骤 4 划分, 直到最后输出最终分类数据和标记.

3.2 融合的聚类诊断策略

诊断系统检测的状态参数的增多,增加了对采样数据聚类的难度,其主要表现在:故障数据状态参数分布空间的复杂化、数据样本的重复冗余和状态参数在维数上的高度冗余.高维数据在处理过程中会产生较大的计算消耗,同时对于这些多属性共同决定的故障问题,很难直观地分析某些主要属性对于系统的影响,从而为找到故障的本质变化和规律,以及在计算机上可视化分析这些数据的结构、关联和分布情况带来了很大困难.本文的 ALLE 与谱聚类融合的聚类算法的具体描述为:①将原始观测空间数据采用 ALLE 算法降维至特征空间;②对特征空间数据递归调用规范切算法进行聚类实验.

4 实验结果与分析

将自适应局部线性嵌入和谱聚类融合方法,分别应用于标准 UCI 数据集^[9]和田纳西-伊斯曼过程的数据故障诊断中,并和常规算法进行对比,以验证本文方法发现故障的能力.相关实验在 CPU 为 Intel Pentium 双核 2.0 GHz,内存为 1 GB 的 PC 机上进行,采用 Matlab 编程实现.

4.1 UCI 数据集降维处理

从 UCI 机器学习数据库^[9]中选取 Iris、Diabetes、Wine、Iris 数据集用萼片和花瓣的长度、宽度来区分 3 种不同的花,数据集共 150 个数据样本,而每一类又分为 50 个数据样本. Diabetes 数据集是对比马印第安人患糖尿病与否进行描述,共 768 个样本,每个样本包括 8 个属性. Wine 数据集记录了葡萄酒的化学分析结果,确定了 3 种葡萄酒每种的 13 个属性的量值,共有 178 个数据样本.

本文分别采用 ALLE、LLE 以及 PCA 算法对 Iris、Wine 数据集分别降至 3 维,并对样本数据在三维空间的分布进行分析,以解决在高维空间无法获得数据点分布图的问题.从图 2 和图 3 中可以看出,3 种算法能够很好地分开第 1 类数据,本文的 ALLE 算法较 LLE 和 PCA 算法对第 2 类和第 3 类数据分类的效果明显. ALLE 算法对于 Wine 数据集除了个别边界点之外,均能很好地区分开,从实验可以看出,ALLE 算法能够较好地发现高维数据在低维空间的内在结构,因此能够正确提取特征信息.根据数据在 3 维空间的分布情况,ALLE 算法具有较好的可分性,同时在低维空间作图,更有利于分析同类数据和类间数据之间的本质联系.

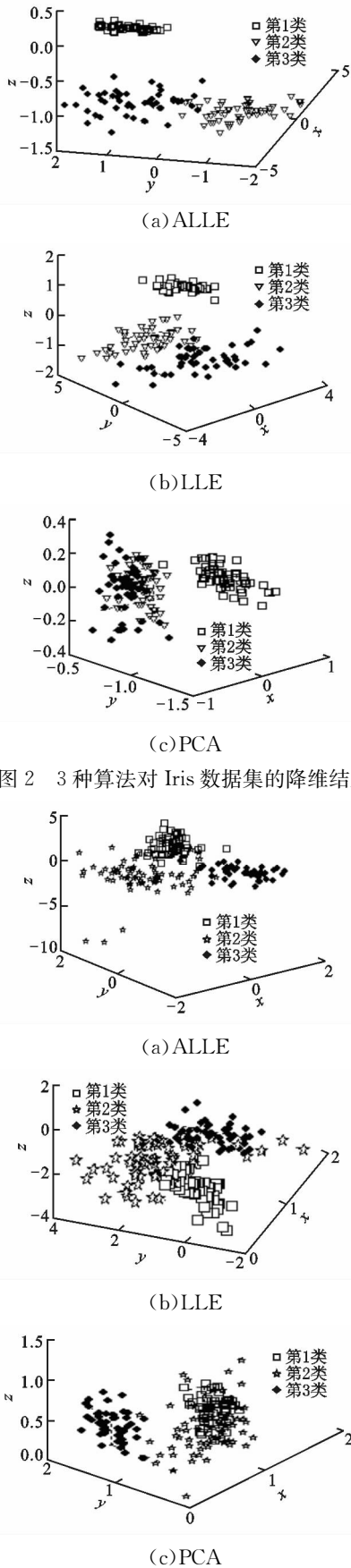


图 2 3 种算法对 Iris 数据集的降维结果

图 3 3 种算法对 Wine 数据集的降维结果

4.2 UCI 数据集对比分类实验

采用 ALLE、LLE 以及 PCA 算法分别将 Iris 和 Diabetes 数据集降至 3 维,Wine 数据集降至 4 维,采用递归调用规范切算法(RANCA)分别对降维前和降维后的 Iris、Diabetes、Wine 数据集进行测试,并验证了对比的准确性和稳定性.实验主要步骤如下:将数据集随机分成 2 份,取其中一份作为训练数据,剩余部分用来验证分类测试效果,对每一种数据集都进行了 50 次独立实验.

递归调用规范切算法的阈值对数据集的分类性能有一定的影响,但影响不大,本文从正确率、标准差方面比较数据集降维前后的分类变化,实验阈值在 0~3 之间变化,每种数据集均进行 50 次实验.应用递归调用规范切算法,对数据集进行降维处理前后的平均正确率 δ 、标准差 σ 的结果如图 4~图 7 所示.

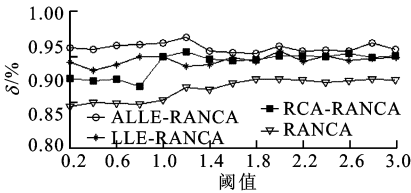


图 4 Iris 数据集在不同阈值下的正确率曲线

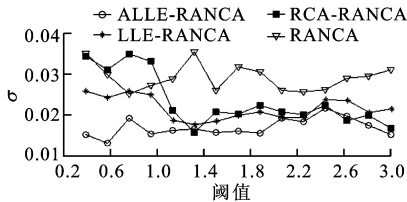


图 5 Iris 数据集在不同阈值下的标准差曲线

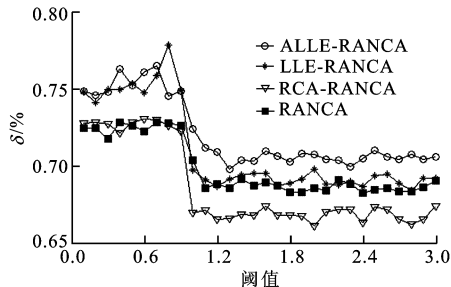


图 6 Diabetes 数据集在不同阈值下的正确率曲线

实验结果表明,对于 Iris、Diabetes 和 Wine 数据集,采用 ALLE 算法降维预处理后,平均分类正确率在不同阈值下均高于原始数据的分类正确率,以及 PCA、LLE 算法预处理数据的分类正确率,同时 σ 较小,稳定性有一定的提高.因此,在低维空间,更有利于对故障本质规律的分析.

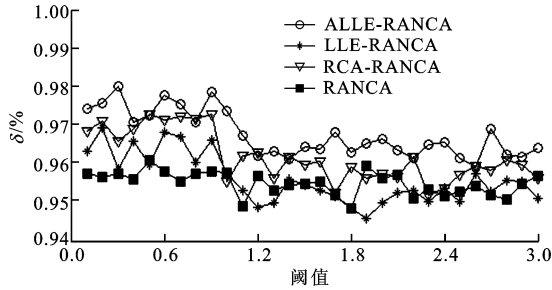


图 7 Wine 数据集在不同阈值下的正确率曲线

通常情况下,随着阈值的增加,递归调用规范切算法提取的特征数据个数相对较多,运算时间也较长,计算时间会在阈值为 1.0 的附近显著增加.但是,经过 ALLE 算法降维处理后的数据集,在阈值小于 1.0 附近时,正确率可达到较大值,说明对降维预处理后的数据进行分类不仅提高了计算速度,而且保证了较高的分类正确率.

4.3 田纳西-伊斯曼过程故障诊断

田纳西-伊斯曼过程(TEP)是由伊斯曼化学公司创建的,其目的是为评价过程控制和监控方法提供一个现实的工业过程.过程包含 5 个主要单元,8 种成分^[10],41 个测量变量和 12 个控制变量.另外,除去反应器搅拌器的搅动速度,还有 52 维观测变量和 22 种状态(正常状态加 21 种故障状态).数据集分为训练数据集和测试数据集,在训练数据集中,每种故障状态取 480 个样本,正常状态取 500 个样本,共 10 580 组样本.测试集中,每组数据包含 960 个样本,其中前 160 个为正常状态,后 800 个为故障状态,共 21 120 组样本.实验的主要步骤为:在训练数据集中提取正常状态和某一种故障状态的样本作为训练数据集,得到数据特征,然后在测试数据集中分别验证这 21 种故障和正常过程的分类正确性.对每一种数据集都进行 50 次独立实验,将数据集分别采用 PCA、模糊人工免疫网络数据分类算法(FAINC)、递归调用规范切(RANCA)、局部线性嵌入与递归调用规范切结合算法(LLE-RANCA),以及本文提出的自适应局部线性嵌入和递归调用规范切融合算法(ALLE-RANCA)进行故障分类实验,部分实验结果见表 1.

对比表 1 中的 5 种算法,除了第 16 号故障诊断正确率略低外,其他均有很好的诊断效果(现有的高维观测属性为 52 维,故障号 16 与正常模式之间,样本重叠严重,所以现有的一些方法的诊断准确率都较低).同时,对于不同故障诊断的稳定性,本文提出的 ALLE-RANCA 融合算法均高于其他算法.

表 1 几种方法的诊断正确率统计

故障号	诊断正确率/%				
	PCA-T ²	FAINC	RANCA	LLE-RANCA	ALLE-RANCA
1	99.20	99.54	100.00	99.65	100.00
5	22.50	47.81	40.54	42.74	51.32
10	33.40	63.71	57.64	60.38	66.57
13	91.50	72.75	81.13	87.42	92.31
16	16.60	61.87	56.55	42.67	54.95
17	74.10	53.47	73.89	74.81	78.21
20	29.90	60.63	80.21	78.64	84.62

5 结 论

在故障诊断过程中,对于高维、非线性数据,很难从观测空间发现嵌入在故障数据中蕴含的规律问题,因此故障诊断包含故障特征的提取和模式识别步骤.本文引入非线性嵌入算法,对高维数据进行维数约简和特征提取,并针对该算法在处理非理想数据时稳定性较差的特点,提出了自适应局部线性嵌入算法,结合递归调用规范切算法,对故障数据进行分类.通过对一些样本进行可视化作图分析,为故障数据提供了低维嵌入空间模式识别的方法.

参考文献:

[1] FRANK P M. Analytical and qualitative model-based fault diagnosis-a survey and some new results [J]. European Journal of Control, 1996, 2(1): 6-28.

[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]. Chinese Mechanical Engineering, 2000, 11(2): 211-216.

[3] 王娜,杜海峰,庄健,等. 用于故障诊断的网络聚类谱分割方法 [J]. 机械工程学报, 2008, 44(10): 228-233.

WANG Na, DU Haifeng, ZHUANG Jian, et al.

Spectral clustering method based on network segmentation used in fault diagnosis [J]. Chinese Mechanical Engineering, 2008, 44(10):228-233.

[4] SEUNG H S, DANIEL D L. The manifold ways of perception [J]. Science, 2000, 290 (12): 2268-2269.

[5] ROWEIS S T, SAUL L K. Nonlinear dimensionality reduction by local linear embedding [J]. Science, 2000, 290: 2323-2326.

[6] SAUL L K, ROWEIS S T. Think globally, fit locally: unsupervised learning of nonlinear manifolds [J]. Journal of Machine Learning Research, 2003, 4(12): 119-155.

[7] SHI J, MALIK J. Normalized cuts and image segmentation [J]. IEEE Transactions on PAMI, 2000, 22(8): 888-905.

[8] GOLUB G H, VANLOAN C F. Matrix computations [M]. Beijing, China: China Machine Press, 2001.

[9] NEWMAN D J, HETTICH S, BLAKE C L. UCI repository of machine learning databases[S/OL]. Irvine, CA: University of California, Department of Information and Computer Science. [2009-04-16]. <http://www.ics.uci.edu/~mllearn/MLRepository.html>.

[10] 蒋浩天,拉塞尔,布拉茨. 工业系统的故障检测与诊断 [M]. 北京: 机械工业出版社, 2003:98-107.

[本刊相关文献连接]

粒计算及其在机械故障智能诊断中的应用. 2009, 43(9):37-41.

具有模糊隶属度的模糊支持向量机算法. 2009, 43(7):40-43.

自适应移频变尺度随机共振在故障诊断中的应用. 2009,43(7):69-73.

气动阀间歇性排气噪声辐射规律的研究. 2008, 42(9):1091-1095.

运用电路蕴涵及拓扑信息提高数字电路故障压缩效率. 2008,42(8):958-962.

(编辑 管咏梅)