

基于慢特征密度聚类的气路异常检测方法

孙昊, 付旭云, 钟诗胜

(哈尔滨工业大学(威海)海洋工程学院, 山东威海264209)

摘 要: 在结合核技巧、慢特征分析算法与密度聚类方法的基础上,提出了基于混合核慢特征分析和密度聚类的慢特征密度聚类算法,实现了基于民航发动机气路参数原始值的异常检测。核技巧的引入克服了慢特征分析法处理复杂数据时可能存在的维度爆炸问题,充分利用不同核函数的特点和慢特征分析的优势从气路参数原始值中提取出随时间变化最缓慢的特征作为密度聚类算法的输入,最终筛选出异常值。经实验对比发现,该方法针对某些异常拥有最好的聚类效果和最低的虚警率,尤其是检测可调放气活门系统异常时虚警数量不到样本总数量的0.5%,是一种有效的方法。

关键词: 民航发动机; 气路; 异常检测; 慢特征分析; 核技巧; 密度聚类

中图分类号: V263.4

文献标志码: A

Slow feature density clustering-based gas path anomaly detection method

SUN Hao, FU Xuyun, ZHONG Shisheng

(School of Ocean Engineering,

Harbin Institute of Technology, Weihai, Weihai Shandong 264209, China)

Abstract: Based on the kernel method, slow feature analysis algorithm and density clustering method, a mixed-kernel slow feature density clustering algorithm was proposed to detect the original gas path parameters of civil aero-engine. The introduction of the kernel method overcame the possible dimensional explosion when the slow feature analysis dealt with complex data, and it took full use of the characteristics of different kernel functions and advantages of slow feature analysis. This algorithm can extract the feature with the slowest time-dependent change from the original gas path parameters and use it as the input of the density clustering algorithm. Then the anomalies were found out. Through experimental validation, this method had the best clustering results and the lowest false alarm rate on certain anomalies. Especially, the number of false alarms was less than 0.5% of the total number of samples when detecting the anomaly of the variable bleed valve system, proving it is an efficient method.

Key words: civil aero-engine; gas path; anomaly detection; slow feature analysis; kernel method; density clustering

收稿日期: 2020-09-11

基金项目: 山东省自然科学基金(ZR2019MEE096); 中央高校基本科研专项资金(HIT.NSRIF.201704);

国家自然科学基金民航联合基金(U1733201)

作者简介: 孙昊(1996—),男,硕士生,主要从事民航发动机气路异常检测。

通信作者: 付旭云(1981—),男,副教授,博士,主要从事智能运维、预测与健康管理研究。E-mail: fuxuyun@hit.edu.cn

引用格式: 孙昊, 付旭云, 钟诗胜. 基于慢特征密度聚类的气路异常检测方法[J]. 航空动力学报, 2021, 36(10): 2218-2229. SUN Hao, FU Xuyun, ZHONG Shisheng. Slow feature density clustering-based gas path anomaly detection method[J]. Journal of Aerospace Power, 2021, 36(10): 2218-2229.

为了提高民航发动机的自主监控能力,各航空公司和科研机构对民航发动机的气路异常检测进行了大量的研究,大致可以分为三类:基于模型的异常检测方法、基于知识的异常检测方法和基于数据的异常检测方法^[1]。目前国内航空公司对航空发动机进行的故障诊断主要依赖发动机制造商提供的客户通知报告(customer notification report, CNR)^[2]。客户通知报告中只提供故障发生的可能原因和气路参数偏差值,不提供具体判断模型和原理,导致独立建立精确的数学或物理模型极其困难,而且客户通知报告也不能保证完全准确的找出全部异常,即客户通知报告的结果不够全面。而基于知识的方法需要建立复杂的知识库,知识库的完备性和可维护性都难以保证。同时,收集带有高质量标签的气路参数数据集较为困难。随着机器学习技术的发展,基于数据驱动的方法可以挖掘数据内部信息,拥有更强的实用性,逐渐广泛的应用于发动机异常检测领域^[3-4]。

根据数据来源的不同,最新的数据驱动的气路异常检测方法主要有以下几种:①通过主成分分析结合深度置信网络、混沌粒子群优化 Elman 神经网络等方法挖掘航空发动机基线^[5-6],再进一步使用粒子群与极限学习机进行后续的异常检测^[7];②通过建立数学模型^[8]、残差前向神经网络^[9]等方法挖掘气路参数偏差值,使用支持向量机进行异常检测^[10];③利用来自快速存取记录器或飞机通信寻址与报告系统的气路参数原始值,通过堆叠去噪自编码器、卷积神经网络^[11]、自编码器结合高斯模型^[12]、多源信息融合结合前馈神经网络^[13]等进行异常检测。

但数据量不足时基线模型难以建立;偏差值挖掘精度有限;气路参数原始值噪声多,关联关系复杂;无论是使用哪种数据进行异常检测任务,都需要首先对数据提取特征,特征提取的好坏直接关系到异常检测的准确性,而正常样本和异常样本的极端不平衡又极大的影响常见异常检测方法的准确性。

针对以上难题,本文提出一种基于慢特征密度聚类的民航发动机气路参数原始值异常检测方法,仅需经过简单的计算、不需要挖掘性能基线和偏差值、对样本标签没有过高需求、不会因为异常样本占比过少影响效果。该方法首先通过结合了核技巧的慢特征分析为气路参数原始值提取特征,既能充分利用高维特征空间的优点,又不增加

计算量,同时,将不同的核函数混合使用,充分利用不同核函数的优点,提取出的特征能够通过明显的突变反映出异常。然后将提取出的特征输入基于密度的聚类算法,利用密度聚类处理不平衡数据集的优势,克服正常和异常样本不平衡导致的检测精度下降的问题。通过相关评价指标评估出的最优聚类结果筛选出多数类(正常类)、少数类和噪声点(异常类),从而实现民航发动机气路参数原始值异常检测。

1 基于混合核慢特征分析的特征提取方法

1.1 慢特征分析算法

慢特征分析(slow feature analysis, SFA)是一种从快速变化的时序信号中抽取随时间变化最缓慢特征的一种算法^[14]。民航发动机气路的性能随时间缓慢变化,正常情况下不会出现剧烈波动,因此可以认为通过慢特征分析算法从民航发动机气路参数原始值中提取出的随时间变化最缓慢的特征一定程度上能够代表反应发动机性能的某些固有特性,根据这些固有特性的变化幅度可以判断发动机的运行状态。

慢特征分析的任务就是找到一个非线性函数 $f(x)$, 输出 $y(t)=f(x(t))$ 随时间变化最缓慢的一组特征。

比如考虑每条数据包含 I 个参数的 I 维时间序列输入:

$$x(t)=[x_1(t), x_2(t), \dots, x_I(t)] \quad (1)$$

式中 $t \in [t_0, t_1]$ 表示时间。

慢特征分析算法的目标是找到如式(2)所示的一个非线性函数,使如式(3)所示的输出随时间变化尽可能缓慢。

$$f(x)=[f_1(x), f_2(x), \dots, f_J(x)] \quad (2)$$

$$y(t)=f(x(t))=[y_1(t), y_2(t), \dots, y_J(t)] \quad (3)$$

该优化问题等价于式(4),同时需要满足式(5)与式(6),使得输出归一化处理并排除无意义的常数解,满足式(7)使得输出间互不相关,从而携带更多信息。

$$\min_{y_j} \Delta(y_j) := \langle (\dot{y}_j)^2 \rangle \quad (4)$$

$$\langle y_j \rangle = 0 \quad (5)$$

$$\langle y_j^2 \rangle = 1 \quad (6)$$

$$\langle y_i y_j \rangle = 0, \quad \forall i < j \quad (7)$$

式中 $\dot{y}_j$ 表示 y_j 关于时间 t 的 1 阶导数; $\langle \cdot \rangle$ 表示对时间 t 求均值。

但 SFA 算法的准确性依赖于非线性函数对原始输入非线性扩展的程度,非线性函数的阶数越高,非线性扩展越充分,算法越能够捕捉到系统本身的固有属性。但随着非线性扩展函数阶数的增加,特征空间维数将指数增加,给算法的计算带来极大的困难,影响 SFA 算法的应用^[15]。

1.2 混合核函数的选择

通过引入核技巧,算法能够获得高维特征空间带来的高精度,同时只需要在原始低维空间中计算,降低了计算复杂度。将核方法引入慢特征分析可以在有限的计算复杂度下极大提高 SFA 对非线性系统的特征提取能力。

根据 Mercer 定理^[16],当核函数 $k(x, y)$ 满足一定条件时,它可以重构为 x, y 的某个非线性映射 ϕ 的内积,如式(8)所示。核方法的计算不需要知道映射 $\phi(x)$ 和 $\phi(y)$ 的具体形式,只是隐式的利用了映射函数及其结果,因此它的计算复杂度只和样本数量有关。

$$k(x, y) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x) \phi_i(y) \quad (8)$$

式中 ϕ 表示映射函数, λ_i 表示线性组合的系数。

假设非线性映射 $\Phi: x \rightarrow \Phi(x), R^m \rightarrow F$ 将原空间 R^m 中的数据 X 映射到高维特征空间 F 中,且算法所需的所有值均可以表示为内积 $k(x, y) = \langle \Phi(x), \Phi(y) \rangle$ 的形式,那么只需要在特征空间中计算核矩阵 K ,利用核矩阵 K 跳过映射函数的具体计算,直接表示映射函数的内积,核矩阵中的元素如式(9)所示。

$$K_{ij} = k(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle \quad (9)$$

式中 x_i 表示第 i 条输入。只要矩阵 K 为半正定矩阵即可满足 Mercer 条件。

对于民航发动机气路参数的特征提取,一部分旧发动机记录了海量的数据,核函数的使用使得不需要知道映射函数和特征空间的具体细节,只需要考虑样本数量带来的计算复杂度,不再需要考虑高维映射函数导致的维数爆炸问题。因此选择恰当的核函数能够快速准确的计算大量气路参数输入数据。

核函数的选择对于算法精度是至关重要的。常见的核函数主要包括高斯核函数、多项式核函数和线性核函数。对于本文研究的问题,使用单一的核函数不能满足精度需要。因此参考相关混

合核函数在支持向量机上的应用,选择以上 3 个核函数的线性组合作为混合核函数^[17]

$$k(x, y) = \lambda \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right) + (1 - \lambda)(x \cdot y)((x \cdot y) + 1)^{q-1} \quad (10)$$

其中 $q \geq 1$, 且为非负整数, $\lambda \in [0, 1]$ 。

根据核函数关于和以及数乘运算封闭的性质可知,多个核函数的线性组合与数乘运算依然是核函数,因此按照上式组合方法得到的函数依然是核函数。

1.3 基于混合核慢特征分析的气路参数原始值特征提取

混合核慢特征分析(mixed-kernel slow feature analysis, MKSFA)的步骤如下所示,首先标准化处理 n 条 m 维输入数据组成的样本数据集 $X_{nm} = [x_1, x_2, \dots, x_n]^T$,本文中 x_i 表示包含 m 个参数的第 i 条样本而不是文献[18]等信号处理领域的 n 条数据的第 i 个参数组成的序列。

设将数据集 X 映射到高维特征空间 F 使用的映射函数为 $\Phi(\cdot)$,则 X 在高维特征空间 F 中的表达式为如式(11)所示:

$$\Theta = [\Phi(x_1), \Phi(x_2), \dots, \Phi(x_n)]^T \quad (11)$$

为去除各维度之间的相关性,简化提取过程,需要对数据进行白化去相关处理。

首先按照式(12)所示方法,对 Θ 中的数据去均值处理,得到去均值处理后数据集 X 在高维特征空间 F 中的表达式 $\tilde{\Theta}$,如式(13)所示。以 $\tilde{\Theta}$ 为基础,按式(14)所示计算特征空间 F 中的协方差矩阵 $\tilde{C}$ 。

$$\tilde{\Phi}(x_j) = \Phi(x_j) - \frac{1}{n} \sum_{i=1}^n \Phi(x_i) \quad (12)$$

$$\tilde{\Theta} = [\tilde{\Phi}(x_1), \tilde{\Phi}(x_2), \dots, \tilde{\Phi}(x_n)]^T \quad (13)$$

$$\tilde{C} = \frac{1}{n} \sum_{i=1}^n \tilde{\Phi}(x_i) \tilde{\Phi}(x_i)^T \quad (14)$$

为进行去相关处理,需要求解式(15)所示的协方差矩阵 $\tilde{C}$ 的特征值 λ 和特征向量 V 。

$$\lambda V = \tilde{C} V \quad (15)$$

V 所有的解都存在于 $\tilde{\Phi}(x_1), \tilde{\Phi}(x_2), \dots, \tilde{\Phi}(x_n)$ 张成的空间中,将 $\tilde{\Phi}(x_i)$ 与式(15)内积后得到式(16)。

$$\lambda \langle \tilde{\Phi}(x_i), V \rangle = \langle \tilde{\Phi}(x_i), \tilde{C} V \rangle \quad (16)$$

式中特征向量 V 为 $\tilde{\Phi}(x_i), i = 1, 2, \dots, n$ 的线性组合,如式(17)所示:

$$V = \sum_{q=1}^n \alpha_q \tilde{\Phi}(x_q) \quad (17)$$

为将核技巧运用在以上白化去相关求解过程中,将 $\Phi(\cdot)$ 替换以降低计算量,分别设核矩阵 K 和标准化核矩阵 $\tilde{K}$ 中的元素如式(18)和式(19)所示。在选择好核函数 $k(x, y)$ 后,不需要具体的映射函数 $\Phi(\cdot)$ 也可将核矩阵 K 和标准化核矩阵 $\tilde{K}$ 中的元素 K_{ij} 和 $\tilde{K}_{ij}$ 视为已知。

$$K_{ij} = \langle \Phi(x_i), \Phi(x_j) \rangle = k(x_i, x_j) \quad (18)$$

$$\tilde{K}_{ij} = \langle \tilde{\Phi}(x_i), \tilde{\Phi}(x_j) \rangle = \tilde{k}(x_i, x_j) \quad (19)$$

则标准化核矩阵 $\tilde{K}$ 中各元素 $\tilde{K}_{ij}$ 与 K 中元素的转换关系如式(20)所示:

$$\tilde{K}_{ij} = \langle \tilde{\Phi}(x_i), \tilde{\Phi}(x_j) \rangle = \frac{(K - 1_n K - K 1_n + 1_n K 1_n)_{ij}}{1_n} \quad (20)$$

式中 1_n 表示 n 阶方阵,其中每个元素都是 $1/n$ 。

将式(14)、式(17)和式(20)带入式(16),化简得到式(21)以简化白化去相关求解过程。

$$n\lambda\alpha = \tilde{K}\alpha \quad (21)$$

式中 α 表示 $\tilde{K}$ 的特征向量, $\alpha = [a_1, a_2, \dots, a_n]^T$; λ 表示前文协方差矩阵 $\tilde{C}$ 的特征值。

求解 $\tilde{K}$ 的全部特征向量,并按特征值由大到小排列选取前 d 个特征向量 $\alpha^1, \alpha^2, \dots, \alpha^d$ 带入式(15)~式(17)中,即可解出协方差矩阵 $\tilde{C}$ 按大小排列的前 d 个特征值 $\lambda_1, \dots, \lambda_d$ 和特征向量 $V^1, \dots, V^d$ 。特征向量个数 d 可以参照经验式(22)选取。

$$\frac{\lambda_i}{\sum(\lambda_i)} > 0.0001 \quad (22)$$

为使特征向量满足 $\langle V^k, V^k \rangle = 1$,需要对 $\alpha^k = [a_1^k, a_2^k, \dots, a_n^k]^T$ 标准化处理,如式(23)所示:

$$1 = \sum_{i,j=1}^n \alpha_i^k \alpha_j^k \langle \tilde{\Phi}(x_i), \tilde{\Phi}(x_j) \rangle = \sum_{i,j=1}^n \tilde{K}_{ij} \alpha_i^k \alpha_j^k = \langle \tilde{K} \alpha^k, \alpha^k \rangle = \lambda_k \langle \alpha^k, \alpha^k \rangle \quad (23)$$

令 $A = [\alpha^1, \alpha^2, \dots, \alpha^d]$, $\Lambda = \text{diag}(\lambda_1 \geq \dots \geq \lambda_d)$,则协方差矩阵 $\tilde{C}$ 的对角化方法如式(24)所示:

$$(\tilde{\Theta}^T \Lambda \Lambda^{-1/2})^T \tilde{C} \tilde{\Theta}^T \Lambda \Lambda^{-1/2} = \Lambda / n \quad (24)$$

如式(25)所示,将样本数据协方差矩阵变换为单位矩阵。

$$(\sqrt{n} \tilde{\Theta}^T \Lambda \Lambda^{-1/2})^T \tilde{C} \sqrt{n} \tilde{\Theta}^T \Lambda \Lambda^{-1} = I \quad (25)$$

则 $\sqrt{n} \tilde{\Theta}^T \Lambda \Lambda^{-1}$ 为白化处理所用的白化矩

阵,通过白化矩阵对特征空间的第 i 条数据 z_i 白化处理结果如式(26)所示。至此,白化处理完成。

$$z_i = \tilde{\Phi}(x_i) \sqrt{n} \tilde{\Theta}^T \Lambda \Lambda^{-1} = \sqrt{n} \tilde{K}_i \Lambda \Lambda^{-1} \quad (26)$$

式中 $\tilde{K}_i$ 表示矩阵 $\tilde{K}$ 第 i 行的所有元素。

则完整数据的白化处理后输出如式(27)所示,且满足 $E(z^T z) = I$ 。

$$z = \sqrt{n} \tilde{K} \Lambda \Lambda^{-1} \quad (27)$$

考虑到样本数量不是很多,慢特征提取部分可以直接采用KPCA(kernel principal component analysis)来处理,对矩阵 $E(\dot{z}^T \dot{z})$ 进行主成分分析,选择它的前 j 个最小特征值对应的特征向量 $w_1, w_2, \dots, w_j$ 作为标准化权重向量。综上所述,最终每个样本点对应的输出如式(28)所示:

$$y_i = z_i W = \sqrt{n} \tilde{K}_i \Lambda \Lambda^{-1} W \quad (28)$$

式中 W 表示标准化权重向量, $W = [w_1, w_2, \dots, w_j]$

则全部输入数据的输出如式(29)所示:

$$Y_{n \times j} = [y_1, y_2, \dots, y_n]^T = z W = \sqrt{n} \tilde{K} \Lambda \Lambda^{-1} W \quad (29)$$

2 基于密度聚类算法的异常检测

聚类(Clustering)是一种无监督学习算法,能够将相似的对象归到同一个簇(类)中^[19]。气路参数提取出的特征在一定范围内小幅波动,正常样本都是相近的,而异常样本与正常样本存在差异。因此将特征提取后的数据输入聚类算法,按照样本分布的紧密程度聚类,游离于所有聚类之外的独立样本和内部样本少的聚类都可以被认为是异常样本,从而实现异常检测的目的。

OPTICS(ordering points to identify the clustering structure)算法(使用排序点识别聚类结构)^[20]是一种基于密度的聚类方法,作为DBSCAN(density-based spatial clustering of applications with noise)算法(具有噪声的基于密度的聚类方法)^[21]的改进算法,将紧密相连的样本划分为同一类从而完成聚类,两者聚类效果相近,但是OPTICS算法对超参数的设置不敏感,少次调参即可得到最佳聚类。

对于气路参数原始值组成的数据集 $X_{n \times m} = [x_1, x_2, \dots, x_n]^T$,使用慢特征密度聚类算法的异常检测流程如下所示:

1) 标准化,首先标准化气路参数原始值,消除量纲对后续计算的影响;

2) 特征提取,设置混合核函数中的超参数,将步骤1)的输出作为混合核慢特征分析的输入进行特征提取;

3) 特征提取效果评估,评估特征提取效果,如果效果不够好,回到步骤2)重新设置超参数,如果效果满足需要,则进行步骤4);

4) 密度聚类,设置初始密度聚类超参数,将特征提取的结果作为输入进行密度聚类;

5) 聚类效果评估,选择多个无监督聚类评价指标评估聚类效果,多次修改步骤4)中的超参数,直到得到最佳聚类;

6) 异常检测,根据聚类结果,选择不属于任何聚类的噪声点和类内样本较少的少数类作为异常;

7) 异常检测准确性评估,检验步骤5)筛选出的异常是否能够找出已知异常,如果找不到,证明算法无效;

8) 虚警率评估,根据实际异常数量、算法找出的异常总量和样本总量评估虚警率,如果虚警率过高,表示算法效果不够好;

9) 如果准确性和虚警率都满足要求,认为本文提出的方法具有实际应用价值。

3 应用验证

为了验证本文提出的基于慢特征密度聚类的异常检测方法能否准确找到已知异常样本,以CFM56-5B-2发动机为例进行应用验证。通常在有足够异常样本的情况下,可以评估方法是否能覆盖全部异常;针对不同异常的特点进行分类,实现故障诊断(fault diagnosis)任务;还可以实现明确的故障等级划分。但是因为当前掌握的数据集中仅有5种异常共6个样本,进行故障等级划分或故障分类等任务会不可避免的存在偶然性,因此目的就是准确的找出这6个异常,进行异常检测任务。

文中第3.1节与3.2节主要以两台可调放气活门(variable bleed valve system,VBV)系统异常的CFM56-5B-2型号发动机为例,通过同样故障原因的两个样本进行详细的应用验证以直观的证明本文方法的优越性,降低偶然性带来的影响。文中第3.3节简单总结全部5种异常的检测结果。

实验考虑包括飞行高度(A_H)、风扇转速百分比(S_f)、核心机转速百分比(S_c)、进口总温(t_a)、排气温度(T_t)、燃油流量(F_t)和马赫数(Ma)在内的

7个监控数据。两台发动机的数据分别为从2012年~2015年间共1190条数据和2009年到2016年间共3830巡航条数据,其中编号为697511的发动机部分数据如表1所示,两台发动机各有一个由CNR报告指出的异常样本。

表1 CFM-5B-2航空发动机气路参数原始值

Table 1 Original gas path parameters of civil aero-engine CFM-5B-2

参数	样本 x_1	样本 x_{557}	样本 x_{1190}
是否异常	否	是	否
航班日期	2012/8/15	2013/11/24	2015/5/12
A_H/m	10 087	10 089	9 175
$S_f/\%$	87.3	88.3	84.8
$S_c/\%$	95.7	96.8	93.5
$t_a/^\circ\text{C}$	-10.2	-11.7	-19.5
$t_c/^\circ\text{C}$	661.1	739.7	609.9
$F_t/(\text{kg/h})$	1 435.287	1 582.419	1 496.342
Ma	0.768	0.77	0.763

首先使用混合核慢特征分析对民航发动机原始气路参数做特征提取,然后使用OPTICS聚类算法对特征进行异常检测,最后找到异常数据。

为了证明算法的有效性,本节在特征提取阶段选择了另外4个不同特征提取方法作为对照;在异常检测阶段在使用了这4个不同特征提取方法进行异常检测的同时,又选择了两种不同的密度聚类算法作为对照。

3.1 基于MKSFA的特征提取

航空发动机原始气路参数存在不同量纲,需要对其标准化处理。为了不改变数据的分布,选用Min-max标准化(也称离差标准化)方法,如式(30)所示:

$$x^* = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (30)$$

式中 $\min(x)$ 表示 x 的最小值, $\max(x)$ 表示 x 的最大值。通过min-max标准化方法,将原始数据线性变换后落到 $[0,1]$ 区间,同时去除量纲。

3.1.1 基于MKSFA的特征提取结果

通过设置合适的超参数,最终得到特征提取结果。经过对提取出的特征作对比,发现虽然提取出的特征互不相关,但他们在发动机气路性能的波动上有相似的体现,过多的增加特征的维度对异常检测精度的提升有限,却增大计算时间。

因此从不断波动的气路参数数据中提取前五个随时间变化最缓慢的特征,这五维特征没有像气路参数原始值的七维数据(比如“燃油流量”和“排气温度”)一样明确的物理意义,但是能够比原始的七维数据更明显的反映发动机气路性能的变化,可以认为其代表某种抽象的气路固有属性,将这一固有属性用于异常检测能获得更好的效果。

如图 1 所示为编号 697511 的发动机经过标准化处理的基于 MKSFA 的特征提取结果,图 2 所示为标准化处理过的气路参数原始值。其中横坐标表示数据点编号,纵坐标数据分布,带有“✕”标记的点表示客户通知报告中给出的已知异常点。

图 1、图 2 中未被标记为“✕”但突出明显的点

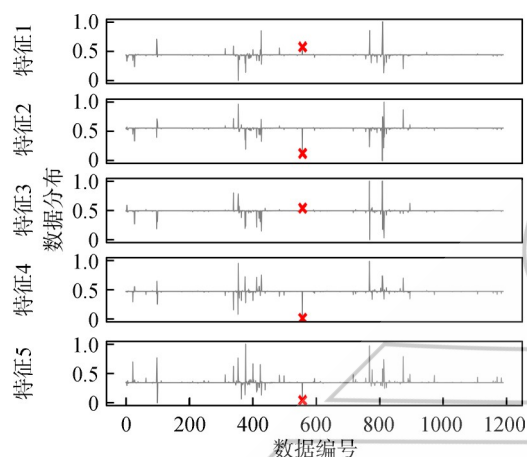


图 1 基于 MKSFA 的特征提取结果

Fig.1 Feature extraction results of MKSFA

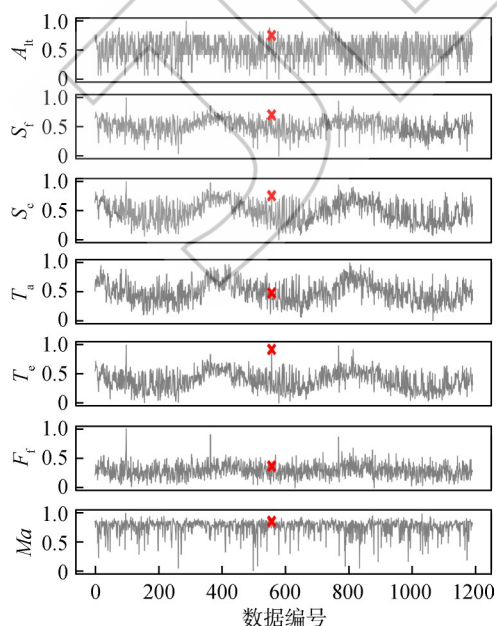


图 2 气路参数原始值

Fig.2 Original gas path parameters

被算法认定与正常值差异较大,并当作异常特征提取出来,这些点有以下几种可能:①可能是未被 CNR 检测到的实际异常(概率较小);②可能是数据超限导致传感器报警等其他未包含在 CNR 中的异常,且传感器报警的数据往往比异常更明显。③可能是 MKSFA 算法错误的将正常数据看作异常即误报(概率较大)。

被超限报警的数据可能由于多种原因导致,未必是发动机真的出现异常,航空公司将直接根据情况处理。CNR 通常用于寻找某些难以直接发现的异常,虽然仍然存在误报和漏报,但准确率较高,CNR 报警后航空公司都会对发动机进行检查。如果能够收集到更高质量的异常标签,将会更好的验证本文方法的有效性。但是 5B 发动机厂家(通用电气)提供的数据接口中仅包含全部的 CNR 报告记录和最近几次的传感器报警。传感器报警数据少,判断异常的准确率不高,因此,本文只能以客户通知报告为准,暂不将这些特征点当作异常特征看待。

根据图 1 和图 2 的对比可以很明显的发现,通过超参数结合不同核函数的优点,使本文提出的特征提取方法具有很高的可靠性,具体表现在:正常数据与异常数据差异明显;携带了异常检测所需要的大量信息,甚至可以通过肉眼判断出明显的异常;提取出的特征互相不相关,具有很强的独立性。因此得出如下结论:基于 MKSFA 的特征提取方法提取的特征能够正确反映航空发动机气路参数的某些固有属性;提取出的正常和异常样本具有明显的不同;相对于原始气路参数,选择合适的 MKSFA 特征提取维度和异常检测算法,能够拥有更好的异常检测效果。

3.1.2 使用单一核函数的特征提取

本节中,分别单独使用高斯核函数、多项式核函数与线性核函数的特征提取结果与使用混合核函数的慢特征提取结果作对比。使用不同核函数的核慢特征(kernel slow feature analysis, KSFA)分析,特征提取结果如图 3~图 4 和图 5 所示,其中横坐标表示数据点编号,纵坐标表示特征提取后数据分布,带有“✕”标记的数据点为客户通知报告给出的已知异常。

根据图 3~图 5 与图 2 的对比,可以很直观的发现将核函数混合后得到的特征提取结果拥有更理想的分布:正常数据分布平稳,异常数据分布较为突出。如表 2 所示为两台发动机使用不同核函

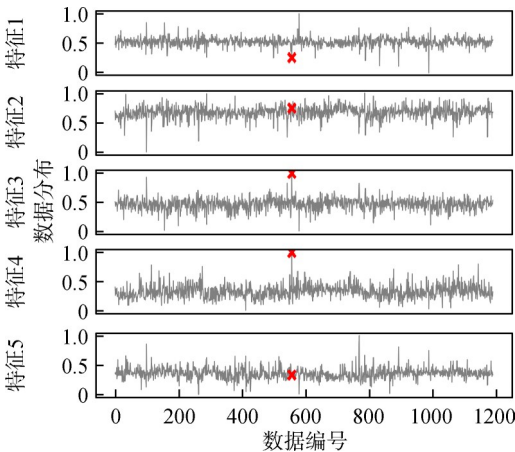


图3 使用高斯核函数的核慢特征提取结果

Fig.3 Feature extraction results of KSFA using Gaussian kernel function

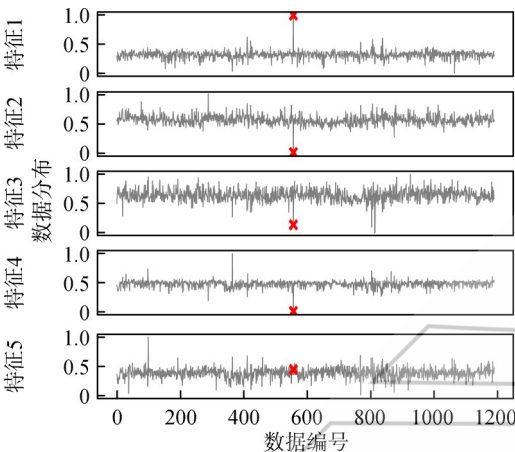


图4 使用多项式核函数的混合核慢特征提取结果

Fig.4 Feature extraction results of KSFA using polynomial kernel function

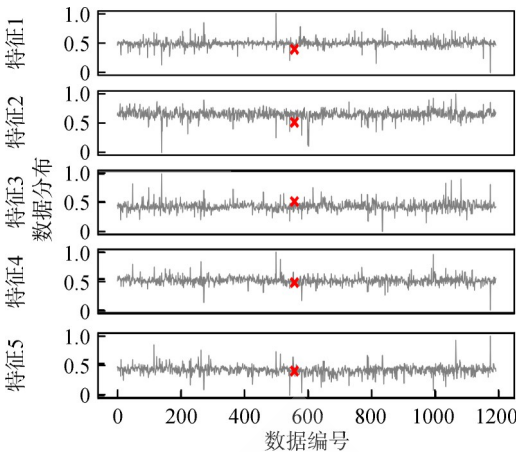


图5 使用线性核函数的混合核慢特征提取结果

Fig.5 Feature extraction results of KSFA using linear kernel function

数得到的特征提取结果的方差,使用混合核函数得到的特征提取结果的方差比单独使用核函数的方差小1~2个数量级。表中加粗字体为各方法中的最优值。

3.1.3 基于MKSFAs的特征提取结果与其他特征提取结果对比

为了证明本文提出的特征提取方法比目前已有的方法效果更好,选择基于主成分分析法(principal component analysis, PCA)^[22]、自编码器(auto-encoder, AE)^[23]、独立成分分析法(independent component correlation algorithm, ICA)^[24]和局部线性嵌入法(locally linear embedding, LLE)^[25]的特征提取结果作对比,对比结果如下所示。

表 2 使用不同核函数的特征提取结果的方差

Table 2 Variances of feature extraction results using different kernel functions

编号	特征	混合核	高斯核	多项式核	线性核
697511	特征 1	0.000 9	0.010 5	0.002 6	0.007 9
	特征 2	0.001 0	0.014 3	0.005 1	0.004 9
	特征 3	0.001 3	0.006 2	0.011 5	0.005 1
	特征 4	0.001 1	0.004 5	0.017 7	0.006 1
	特征 5	0.001 3	0.017 0	0.012 2	0.007 4
699640	特征 1	0.000 2	0.004 9	0.003 1	0.002 1
	特征 2	0.000 4	0.009 4	0.009 0	0.001 0
	特征 3	0.000 4	0.003 3	0.013 0	0.001 3
	特征 4	0.000 4	0.008 8	0.004 2	0.003 0
	特征 5	0.000 2	0.004 0	0.003 6	0.001 7

如图 6~图 9 所示为编号为 697511 的发动机分别基于 PCA、AE、ICA 和 LLE 的特征提取结果。其中横坐标表示数据点编号,纵坐标表示特征提取后数据分布,带有“×”标记的点代表 CNR 报告中给出的异常点。

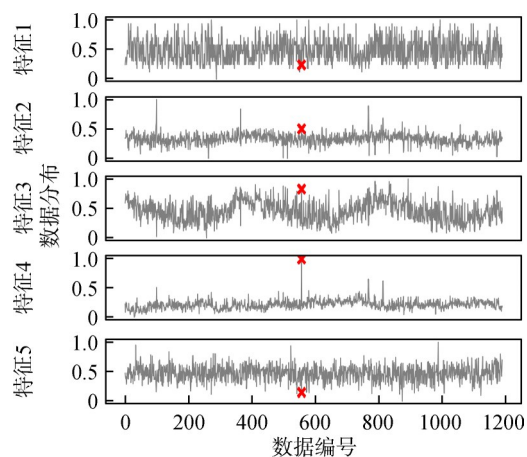


图 6 基于 PCA 的特征提取结果

Fig.6 Feature extraction results of PCA

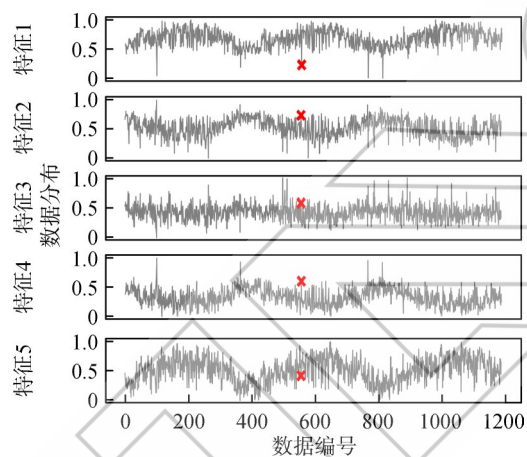


图 7 基于 AE 的特征提取结果

Fig.7 Feature extraction results of AE

如表 3 所示为两台发动机使用不同算法的特征提取结果的方差,本文的特征提取方法提取出的特征相对于其他特征提取方法提取出的特征,正常数据更为集中,异常数据可以肉眼区分。697511 号发动机特征提取结果的方差比其他方法的方差小一个数量级,699640 号发动机由于数据量更大,慢特征提取结果更为充分,得到的方差比其他方法小两个数量级。

因为航空发动机气路参数是高度相关的多维时间序列,不同参数间有复杂的联系,因此以自编码器和主成分分析法为代表的传统特征提取方法难以捕捉到参数中随时间变化的特征,特征提取

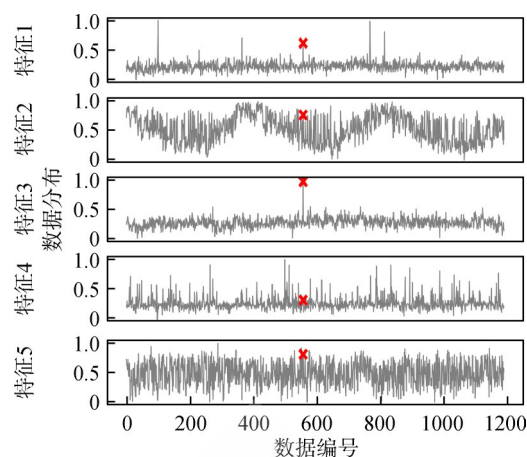


图 8 基于 ICA 的特征提取结果

Fig.8 Feature extraction results of ICA

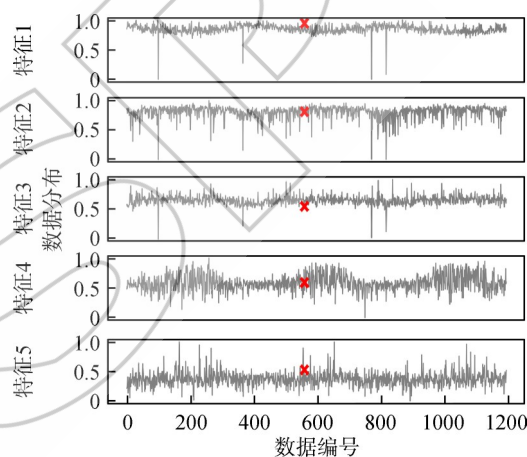


图 9 基于 LLE 的特征提取结果

Fig.9 Feature extraction results of LLE

过程中会损失大量必要信息,导致结果波动很大。根据不同特征提取算法的结果图和方差可以认为本文提出的基于混合核慢特征分析的特征提取方法将会在异常检测领域拥有更好的效果。

3.2 基于慢特征密度聚类的民航发动机异常检测

3.2.1 基于慢特征密度聚类的气路异常检测

如表 4 所示为不同算法在这几个评价指标下的性能对比。为评估无监督聚类算法的效果,引入以下几个评价指标:紧密性(C_p)、间隔性(S_p)、CH (Calinski - Harabasz) 系数(C_h)^[26]、轮廓系数(S_c)^[27]、戴维森堡丁指数(I_{db})^[28]和邓恩指数(I_{dv})^[29]。其中紧密性越小的同时间隔性越大越好;CH 系数和邓恩指数越大越好;轮廓系数越接近 1 越好;戴维森堡丁指数越接近 0 越好。

为评估聚类结果在气路异常检测领域的应用效果,本文对虚警数量和虚警率做出如下定义:定义虚警数量为算法找出的异常样本数量减去已知

表 3 不同特征提取方法的方差

Table 3 Variances of different feature extraction methods

编号	特征	MKSFA	PCA	AE	ICA	LLE
697511	特征 1	0.000 9	0.037 0	0.036 7	0.006 0	0.005 2
	特征 2	0.001 0	0.008 3	0.043 8	0.053 3	0.011 6
	特征 3	0.001 3	0.031 1	0.036 1	0.007 0	0.005 6
	特征 4	0.001 1	0.005 5	0.017 0	0.012 1	0.018 0
	特征 5	0.001 3	0.016 8	0.014 6	0.040 3	0.014 1
699640	特征 1	0.000 2	0.026 8	0.020 4	0.008 3	0.016 7
	特征 2	0.000 4	0.006 0	0.033 3	0.044 0	0.009 9
	特征 3	0.000 4	0.016 7	0.024 8	0.024 5	0.010 8
	特征 4	0.000 4	0.010 4	0.016 1	0.014 3	0.005 7
	特征 5	0.000 2	0.003 4	0.021 1	0.004 7	0.004 8

表 4 使用不同特征提取方法的聚类性能对比

Table 4 Comparison of clustering performance using different feature extraction methods

发动机编号	提取方法	C_p	S_p	C_h	S_c	I_{db}	I_{dv}	R_{fa}
697511	原始数据	0.356	0.583	7.97	-0.427	1.516	0.02	15.2%
	MKSFA	0.013	0.802	149.063	0.979	1.884	0.102	0.5%
	PCA	0.269	0.453	5.851	-0.383	1.379	0.026	21.4%
	AE	0.337	0.561	4.13	-0.452	1.423	0.009	15.5%
	ICA	0.302	0.511	4.772	-0.457	1.409	0.018	无效
	LLE	0.196	0.707	15.968	0.06	1.705	0.013	无效
699640	原始数据	0.012	0.232	72.991	0.952	5.602	0.015	0.4%
	MKSFA	0.012	0.500	460.635	0.983	1.686	0.573	0.2%
	PCA	0.232	0.506	10.280	-0.082	1.528	0.033	3.1%
	AE	0.319	0.530	3.110	-0.200	2.809	0.040	2.5%
	ICA	0.287	0.536	4.934	-0.362	1.388	0.037	无效
	LLE	0.194	0.547	13.847	0.162	1.463	0.014	无效

异常数量;考虑到已知异常数量相同的情况下,样本总量越多,对虚警数量的容忍程度必然越高,因此定义虚警率(R_{fa})为虚警数量与样本总量的比值。

根据表 4 所示结果,基于原始值的密度聚类效果很差,虽然能够找出异常,但是虚警率较高;基于 PCA 和 AE 的聚类效果较差,虚警率较高;基于 ICA 和 LLE 的聚类不能准确找到异常;除戴维森堡丁指数以外,各指标均为基于 MKSFA 的密度聚类方法效果最好,保证了异常检测正确率且降低了虚警率。

3.2.2 基于其他密度聚类算法的气路异常检测

基于密度峰值的聚类算法(clustering by fast search and find of density peaks, CFSFDP)^[30]能够自动地发现簇中心,实现任意形状数据的高效聚类。该算法最大的优点是速度快,不需要迭代。

如表 5 所示为特征提取后数据分别使用 CFSFDP 算法和 OPTICS 算法的结果对比。虽然 CFSFDP 算法速度较快,但是其聚类效果相比于 OPTICS 算法并没有优势,除戴维森堡丁指数略好以外,其余指标均存在明显不足,因此 OPTICS 算法更适合本文提出的方法。

表 5 两种密度聚类算法的聚类性能对比
Table 5 Comparison of clustering performance using two density clustering methods

评价 指标	发动机 697511		发动机 699640	
	OPTICS	CFSFDP	OPTICS	CFSFDP
C_p	0.013	0.011	0.012	0.009
S_p	0.802	0.101	0.5	0.117
C_h	149.063	104.449	460.635	52.917
S_c	0.979	0.907	0.983	0.726
I_{db}	1.884	1.616	1.686	1.426
I_{dv}	0.102	0.002	0.573	0.002
$R_{in}/\%$	0.5	21	0.2	17.5

3.3 其余异常的检测效果

除 VBV 系统异常的两个样本外,实验还使用了 4 台分别发生高压涡轮磨损 (high pressure turbine distress, HPTD)、高压涡轮叶片磨损 (high pressure turbine blade distress, HPTBD)、出口导流片密封叶片缺失 (missing ogv leaf seal, mols) 和排气温度指示异常 (exhaust gas temperature indication, EGT) 故障的发动机验证算法针对不同异常的检测能力。由于这 4 种异常各含有一个样本。为了避免偶然性,本文并未进一步进行故障等级划分或故障诊断即故障分类任务,只进行了找出异常的“异常检测”任务。如表 6 所示为基于不同特征提取方法的密度聚类算法检测以上 4 种故障的虚警率。

这些异常分别具有不同的特点,但是大部分

表 6 基于不同特征提取方法的密度聚类异常检测虚警率
Table 6 FAR of density clustering anomaly detection based on different feature extraction methods

故障类型	特征提取方法	$R_{in}/\%$
HPTBD	MKSFA	3.5
	其余方法	无效
MOLS	MKSFA	5.0
	其余方法	无效
HPTD	MKSFA	16.7
	其余方法	无效
EGT	MKSFA	无效
	其他方法	无效

异常都能体现在气路参数随时间变化最缓慢的某种抽象特征即“慢特征”中,因此可以用密度聚类实现异常检测。这些异常无法明显体现在 PCA、AE、ICA 和 LLE 的特征提取结果中,因此无法使用这 4 种方法和密度聚类实现以上异常的检测。

由于异常样本过少带来的偶然性,初步判定本文提出的方法对于 EGT 指示异常无效,能检测出 6 个异常样本中的其余 5 个,且有最高的准确性和最好的检测效果。综上所述,可以认为本文提出的慢特征密度聚类算法在民航发动机气路异常检测领域拥有良好的表现,是一种拥有广泛应用前景的方法。

3.4 结果分析

对于不同异常类型的检测效果不同的问题,经过对气路参数原始值的分析初步认为:可调放气活门系统异常对发动机某些固有特性影响较大,异常引起提取出的多个慢特征剧烈突变,故异常检测效果较好;高压涡轮叶片磨损和出口导流片密封叶片缺失等异常对某些固有特性影响较小,在慢特征上表现为不是十分剧烈的波动,因此异常检测效果稍差;针对其他无法准确检测到的异常,初步认为它们不会在慢特征上表现出剧烈突变,因而不能通过慢特征对其进行特征提取。6 台发动机提取出的特征中,也并不能反映出不同异常的不同特点,因此无法进行故障诊断任务。

航空发动机是高可靠性设备,异常概率低,但是异常后果严重,不同异常有不同特点,没有一种算法能够全面的检测出全部异常。航空公司将会交叉使用多种不同方法,从而减少漏报,因此通常航空公司所用算法的虚警即误报较多,且难以降低。通常衡量一个异常检测方法的好坏主要靠两个指标:尽可能少的漏报和误报,但由于漏报和误报的矛盾性,很难有一种算法既不漏报也不误报。表 7 所示为不同异常检测方法性能对比结果,多组实验分别采用了不同的方法,可以看出本文提出的气路异常检测方法能够找到已知 6 个异常样本中的 5 个,覆盖了大部分的已知异常,仅漏报一个,总 R_{in} 为 3.66%,小于航空公司要求的误报上限,能够满足要求。因此本文的方法能够满足航空公司的实际需求。

因此可以得出如下结论:①混合核函数降低计算复杂度的同时,充分利用了不同核函数的特点,提取出的慢特征能够反映出异常的差异。②结合 OPTICS 密度聚类算法得到的异常检测结果能够

表 7 不同方法的异常检测结果对比
Table 7 Comparison of anomaly detection using different methods

方法	可检测到的异常覆盖率	检出异常数/总样本数	$R_{fa}/\%$
本文	5/6	462/12 623	3. 66
原始数据+OPTICS	1/6	196/5 020	3. 90
原始数据+CFSFDP	1/6	920/5 020	18. 33
PCA+OPTICS	1/6	292/5 020	5. 82
PCA+CFSFDP	1/6	341/5 020	6. 79
AE+OPTICS	1/6	612/5 020	12. 19
其他	0/6		

更准确的找出异常。③本文方法漏报和误报少，是一种具有广泛应用前景的方法。

4 结 论

针对民航发动机气路参数原始值异常检测的需要,本文提出一种使用混合核慢特征分析法特征提取和使用 OPTICS 密度聚类算法异常检测的慢特征密度聚类异常检测方法。经实验检验,本章提出的特征提取方法,拥有目前最好的特征提取效果。OPTICS 聚类算法用于对提取出的特征异常检测,拥有几乎最高的聚类精度和最低的虚警率。通过实验验证,总结如下:

- 1) 本文使用的混合核函数能够充分利用不同核函数的优点;
- 2) 将慢特征分析法与混合核函数相结合,提高精度的同时减小计算量,提取出的随时间变化最缓慢的几个特征能够明显反映出气路某些抽象固有特性的突变;
- 3) 使用 OPTICS 密度聚类算法进行聚类,聚类精度最高的同时,虚警率最低,OPTICS 密度聚类是最适合本文异常检测任务的聚类方法;
- 4) 本文方法简单方便,有效解决了民航发动机气路异常检测的难题,具有一定的实际应用价值。

后续将会继续研究算法的优化方法;收集更多数据,确定不同种类异常的检测效果;最后重点研究某些异常无法体现在慢特征上的原因。

参考文献:

[1] 钟诗胜,李旭,张永健.基于 DBN 的不均衡样本驱动民航发动机故障诊断[J].航空动力学报,2019,34(3):708-716.
ZHONG Shisheng, LI Xu, ZHANG Yongjian. Fault diagnosis of civil aero-engine driven by unbalance samples based on

DBN[J].Journal of Aerospace Power, 2019, 34(3):708-716. (in Chinese)
[2] QU J F, JIANG S Q, YUAN S H, et al. Analysis of an aero-engine gas path fault monitoring sensor model[J]. Advanced Materials Research, 2014, 1016:516-520.
[3] 付松.基于深度特征挖掘的民航发动机气路故障诊断技术研究[D].哈尔滨:哈尔滨工业大学,2018.
FU Song. Research on gas path fault diagnosis technology for civil aero-engine based on deep feature mining[D]. Harbin: Harbin Institute of Technology, 2018. (in Chinese)
[4] 杨洪富,贾晓亮,任寿伟.基于数据驱动的航空发动机故障诊断与预测方法综述[J].航空精密制造技术,2016,52(5):6-9.
YANG Hongfu, JIA Xiaoliang, REN Shouwei. Review of data-driven aeroengine fault diagnosis and prognosis methods [J]. Aviation Precision Manufacturing Technology, 2016, 52(5): 6-9. (in Chinese)
[5] 王奕首,余映红,卿新林,等.基于 KPCA 和 DBN 的航空发动机排气温度基线模型[J].航空发动机,2020,46(1):54-60.
WANG Yishou, YU Yinghong, QING Xinlin, et al. Exhaust gas temperature baseline model of aeroengine based on kernel principal component analysis and deep belief network[J]. Aeroengine, 2020, 46(1):54-60. (in Chinese)
[6] 瞿红春,林文斌,许旺山,等.基于混沌 PSO_Elman 网络的航空发动机基线挖掘[J].中国民航大学学报,2020,38(1):19-23,53.
QU Hongchun, LIN Wenbin, XU Wangshan, et al. Aero-engine baseline mining based on chaotic PSO_Elman network [J]. Journal of Civil Aviation University of China, 2020, 38(1): 19-23, 53. (in Chinese)
[7] 林文斌.基于数据和知识驱动的发动机监控技术的研究[D].天津:中国民航大学,2019.
LIN Wenbin. Research on engine monitoring technology driven by data and knowledge[D]. Tianjin: Civil Aviation University of China, 2019. (in Chinese)
[8] 崔智全.民航发动机气路参数偏差值挖掘方法及其应用研究[D].哈尔滨:哈尔滨工业大学,2013.
CUI Zhiqian. Civil aeroengine gas path parameter deviation

- mining method with application[D]. Harbin: Harbin Institute of Technology, 2013. (in Chinese)
- [9] ZHOU Xingjie, FU Xuyun, ZHAO Minghang, et al. Regression model for civil aero-engine gas path parameter deviation based on deep domain-adaptation with Res-BP neural network[J]. Chinese Journal of Aeronautics, 2021, 34(1):79-90.
- [10] 刘翠芳, 冯学晓. 小型多旋翼无人机发动机气路故障残差参数偏差值挖掘[J]. 科技通报, 2016, 32(10):184-187.
- LIU Cuifang, FENG Xuexiao. Mining of gas path fault residuals parameter deviation for small uav rotorcraft engine[J]. Bulletin of Science and Technology, 2016, 32(10):184-187. (in Chinese)
- [11] LUO H, ZHONG S. Gas turbine engine gas path anomaly detection using deep learning with Gaussian distribution[C]// Proceedings of the 2017 Prognostics and System Health Management Conference (PHM-Harbin). New York: Institute of Electrical and Electronics Engineers, 2017:1-6.
- [12] FU Xuyun, CHEN Haibo, ZHANG Guangyao, et al. A new point anomaly detection method about aero engine based on deep learning[C]// 2018 International Conference on Sensing, Diagnostics, Prognostics, and Control. New York: Institute of Electrical and Electronics Engineers, 2018:176-181.
- [13] 龚钊宇. 基于人工神经网络的民航发动机故障诊断技术[D]. 成都: 电子科技大学, 2017.
- GONG Zhaoyu. Fault diagnose technology of civil aero engine based on artificial neural network[D]. Chengdu: University of Electronic Science and Technology of China, 2017. (in Chinese)
- [14] WISKOTT L, BERKES P, FRANZIUS M, et al. (2011) Slow feature analysis[EB/OL]. [2020-09-05]. http://scholarpedia.org/article/Slow_feature_analysis.
- [15] 马奎俊, 韩彦军, 陶卿, 等. 基于核的慢特征分析算法[J]. 模式识别与人工智能, 2011, 24(2):153-159.
- MA Kuijun, HAN Yanjun, TAO Qing, et al. Kernel-based slow feature analysis[J]. Pattern Recognition and Artificial Intelligence, 2011, 24(2):153-159. (in Chinese)
- [16] COURANT R, HILBERT D. Methods of mathematical physics: partial differential equations[M]. New Jersey, USA: John Wiley and Sons, 2008.
- [17] 郑慢慢. 基于支持向量机的混合核函数研究[D]. 上海: 华东师范大学, 2019.
- ZHENG Manman. Research on mixed kernel function based on support vector machine[D]. Shanghai: East China Normal University, 2019. (in Chinese)
- [18] 卢依容. 基于核慢特征分析算法的故障检测与诊断[D]. 上海: 上海交通大学, 2015.
- LU Yirong. Fault detection and diagnosis based on kernel slow feature analysis[D]. Shanghai: Shanghai Jiao Tong University, 2015. (in Chinese)
- [19] HARRINGTON P. 机器学习实战[M]. 李锐, 李鹏, 曲亚东, 译. 北京: 人民邮电出版社, 2013.
- [20] ANKERST M, BREUNIG M M, KRIEGER H, et al. OPTICS: ordering points to identify the clustering structure[J]. International Conference on Management of Data, 1999, 28(2):49-60.
- [21] SCHUBERT E, SANDER J, ESTER M, et al. DBSCAN revisited: why and how you should (still) use DBSCAN[J]. ACM Transactions on Database Systems (TODS), 2017, 42(3):1-21.
- [22] WOLD S, ESBENSEN K, GELADI P. Principal component analysis[J]. Chemometrics and Intelligent Laboratory Systems, 1987, 2(1/3):37-52.
- [23] GOODFELLOW I, BENGIO Y, COURVILLE A. Deep learning[M]. Cambridge, US: MIT Press, 2016.
- [24] COMON P. Independent component analysis, a new concept? [J]. Signal Processing, 1994, 36(3):287-314.
- [25] ROWEIS S T, SAUL L K. Nonlinear dimensionality reduction by locally linear embedding[J]. Science, 2000, 290(5500):2323-2326.
- [26] CALIŃSKI T, HARABASZ J. A dendrite method for cluster analysis[J]. Communications in Statistics-theory and Methods, 1974, 3(1):1-27.
- [27] ROUSSEEUW P J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis[J]. Journal of Computational and Applied Mathematics, 1987, 20:53-65.
- [28] DAVIES D L, BOULDIN D W. A cluster separation measure[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1979(2):224-227.
- [29] DUNN J C. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters[J]. Journal of Statistical Computation and Simulation, 1973, 91(13):2699-2716.
- [30] RODRIGUEZ A, LAIO A. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191):1492-1496.

(编辑: 张雪)