DOI:10.3724/zrht.1674-5825.2025103

置换特征重要性机制下空间应用热控系统可解释的在线故障检测与定位方法

张竞菲[1], 王红飞[1], 王亦风[1,2*], 张陈晨[1], 宋 磊[1], 金 山[1], 郭晓晓[1]
(1.中国科学院空间应用工程与技术中心,北京 100094;2.南京航空航天大学计算机科学与技术学院,南京 211106)

摘 要:针对空间应用热控系统(SA-TCS)开展故障诊断时需要同时应对数据流概念漂移的问题与在线处理实时性的要求,基于传统静态模型的故障诊断方法难以适用。为此,构建了一种基于增量置换特征重要性(iPFI)的可解释性在线故障诊断框架。该框架利用增量学习模型实时预测系统的关键参数和健康状态,并利用iPFI算法量化各传感器采样特征对于模型预测的全局重要性。通过监测特征重要性突变实现了故障报警信息的双向验证,并实时定位指示故障的关键传感器及其关联的系统部件。通过模拟多工况SA-TCS的管道泄漏及部件失效故障生成了多工况故障数据集,并以数据流的形式验证了所提出的在线故障检测与定位方法的有效性和优势。实验结果表明:所构建的模型可准确捕捉由于工况突变和故障事件导致的特征重要性动态变化,实现了多工况SA-TCS准确、实时的故障检测与定位。

关键词:故障检测;故障定位;热控系统;空间应用;概念漂移;可解释机器学习;置换特征重要性
中图分类号:V11; TP183 **文献标识码**:A **文章编号**:1674-5825(2026)01-0039-13

Interpretable Online Fault Detection and Localization for Space Thermal Control Systems Using Permutation Feature Importance

ZHANG Jingfei[1], WANG Hongfei[1], WANG Yifeng[1,2*], ZHANG Chenchen[1], SONG Lei[1], JIN Shan[1], GUO Xiaoxiao[1]
(1. Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing 100094, China;
2. College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)

Abstract: Thermal control systems for space applications (SA-TCS) pose significant challenges for fault diagnosis, as they must simultaneously cope with concept drift in data streams and real-time processing requirements. Traditional fault diagnosis methods based on static models are often inadequate handling concept drift and real-time diagnostic demands. To overcome these limitations, an interpretable online fault diagnosis framework based on incremental permutation feature importance (iPFI) is proposed. The framework employs an incremental learning model to predict key system parameters and health status of the SA-TCS in real time, while using the iPFI algorithm to dynamically track the global importance of each sensor feature

收稿日期:2025-12-12; 修回日期:2026-01-21
基金项目:国家资助博士后研究人员计划(GZB20240766); 中国博士后科学基金第77批面上资助(2025M771732); 国家自然科学基金青年科学基金项目(C类)(62503453); 中国科学院空间应用工程与技术中心空间应用创新行动课题(CSU-CXXD-PY-2025-001, CSU-CXXD-ZD-2025-001); 国家重大科技专项空间站运营期载荷运管工作(T3142411SN)
第一作者:张竞菲, 女, 博士, 助理研究员, 研究方向为空间站有效载荷的智能监控与故障诊断技术。E-mail:zhangjf@csu.ac.cn
* 通讯作者:王亦风, 男, 硕士, 高级工程师, 研究方向为有效载荷软件设计及测试、天地协同智能运控技术研究与应用。E-mail:wangyifeng@csu.ac.cn
引用格式:张竞菲,王红飞,王亦风,等. 置换特征重要性机制下空间应用热控系统可解释的在线故障检测与定位方法[J]. 载人航天,2026,32(1):39-51.
Zhang J F,Wang H F,Wang Y F,et al. Interpretable online fault detection and localization for space thermal control systems using permutation feature importance[J]. Manned Spaceflight,2026,32(1):39-51.

with respect to the model predictions. By monitoring abrupt changes in feature importance, fault alerts are bidirectionally verified, and both the key sensors and their associated system components indicating faults are identified in real time. This study simulates pipeline leakage and component failure scenarios under multiple operating conditions of the SA-TCS. The effectiveness and advantages of the proposed online fault diagnosis framework are validated using the simulated dataset. Experimental results show that the constructed model can accurately capture dynamic changes in feature importance caused by condition shifts and fault events, thereby achieving accurate and real-time fault detection and localization for multi-condition SA-TCS.

Key words: fault detection; fault localization; Thermal Control System (TCS); space applications; concept drift; Explainable Artificial Intelligence (XAI); Permutation Feature Importance (PFI)

1 引言

空间应用热控系统(Thermal Control System for Space Applications,SA-TCSs)的稳定运行对于保障在轨设备性能和航天员生命安全至关重要。作为一种典型的复杂机电系统,SA-TCS的运行状态受到任务指令、负载变化和航天器平台调控等多重因素影响,呈现出鲜明的多工况特性[1]。SA-TCS在不同运行工况模式间的切换,将导致传感器监测数据(如温度、压力、流量等)的统计分布和变量间的内在关联性发生动态演化,这一现象在机器学习领域中被称为概念漂移[2]。因此,SA-TCS非平稳的监控数据特性使建立高精度且低误报的系统状态监控与故障诊断模型极具挑战性。事实上,若未能及时、准确地发现异常或故障,可能引发连锁反应,甚至造成灾难性后果。在此背景下,迫切需要面向在轨SA-TCS建立可在线、自适应学习的智能故障诊断框架[3,4]。

现有的故障诊断方法可归类为基于机理模型与数据驱动两类[5]。机理模型的故障诊断方法基于热力学方程和流体动力学方程建立系统精确的物理模型,以实现对系统异常状态的诊断。然而,系统未知的参数和扰动信息,以及未建模故障动态可能对其诊断效能和泛化性能产生较大影响[6,7]。数据驱动的故障诊断方法一般通过统计过程监控和智能学习的方法来实现。基于统计的方法不依赖于对系统复杂机理过程的建模,但其诊断性能一般建立在监测数据平稳性或线性的假设之上,因而难以有效捕捉SA-TCS监控数据的高度非线性与时变特性[8]。近年来,人工智能技术飞速发展,深度学习模型为解决此类问题提供了新范式。深度神经网络具有从高维数据中自学习复杂非线性表征的能力,在故障分类[9,10]、时序依赖性挖掘[11]和拓扑结构信息利用[12,13]等任务上展现出了卓越性能。

尽管上述智能故障诊断方法已取得了显著进展,但在应用于SA-TCS这类强实时性数据流的监控系统时,仍面临着以下三方面的严峻挑战。①动态适应性匮乏:多数智能诊断模型采用离线批处理范式进行训练,导致其部署后模型结构固化,难以及时响应由工况变化引起的概念漂移[2];②概念漂移与真实故障的解耦困难:现有在线学习方法将工况切换与异常事件引发的数据分布偏移统一视为概念漂移,尚缺乏有效的解释机制,因而难以对引起数据流概念漂移的正常操作与真实的异常或故障进行区分;③精细故障诊断算法的通用性欠佳:现有数据特征提取及故障定位算法往往针对特定系统特点而设计,难以直接迁移至结构复杂、监测变量繁多的SA-TCS等系统。目前,上述问题严重制约了现有在线学习方法在实际多工况复杂系统过程监控场景中的广泛应用。

实际上,现有智能故障诊断研究大多侧重于精准完成故障检测和识别任务。然而,当模型进行故障报警时,难以追溯异常或故障事件的根本原因[14],也无法为运维人员提供清晰的决策依据。在此情况下,由于深度学习模型判别结果的可信性不佳,现有智能故障诊断方法在应用于在轨SA-TCS时无法满足其高可靠性的要求。由于智能诊断模型尚无法自动定位触发报警的具体传感器或部件,在轨异常和故障的快速响应与决策也面临着巨大挑战。近年来,可解释人工智能(Explainable Artificial Intelligence,XAI)为解决上述模型可信性问题提供了新范式[15]。其中,置换特征重要性(Permutation Feature Importance,PFI)是一种典型的与模型无关的全局解释方法[16]。PFI通过随机重排单个特征的取值并观察模型性能的下降程度来衡量该特征的重要性。然而,PFI算法需要针对完整数据集进行运算,不适用于数据流和在线学习

场景。为此，Fumagalli等[17]提出了增量形式的置换特征重要性(Incremental Permutation Feature Importance，iPFI)算法，在每个时刻估计单样本重要性，并聚合历史信息，实现了对PFI的在线、低复杂度近似，能够即时响应概念漂移，生成与数据流并行的解释信息流。尽管如此，如何将上述解释器推广至故障诊断框架中，并利用其解释信息实现故障的精细分析，仍然是一个亟待研究的科学问题。

本文针对多工况SA-TCS的故障检测与定位问题展开研究，构建了一种集成增量学习模型与iPFI解释器的在线故障诊断新框架。通过将iPFI方法引入在线故障检测任务，利用其对概念漂移的快速响应和在线解释能力获得特征重要性的动态变化，辅助区分工况切换与突发故障或异常事件，进而基于特征解释信息流实时定位与故障强相关的传感器监测变量，以实现从故障检测到故障定位的贯通。

2　问题的形式化描述

考虑一个具有多传感器采样的动态系统，则t ($t=1,2,\cdots,T$)时刻的实时数据可表示为$\{(\boldsymbol{x}_t,\boldsymbol{y}_t)\}$，其中，$\boldsymbol{x}_t$表示$t$时刻的传感器读数向量(特征)，$\boldsymbol{y}_t$表示$t$时刻系统的健康状态标签。在实际应用中，$\boldsymbol{y}_t$通过专家标注或基于某些性能指标间接获得。进行故障检测或健康状态识别的标签空间通过$\boldsymbol{Y}=\{F_0,F_1,\cdots,F_m\}$表示，其中$F_0$表示正常运行状态，$F_1$~$F_m$表示$m$种故障状态。特别地，在故障检测任务中，满足$m=1$；在故障分类(健康状态识别)任务中，满足$m>1$。在上述标签空间下，$\boldsymbol{y}_t\in\{0,1\}^{|\boldsymbol{Y}|}$且$\sum_{\gamma=1}^{m+1}\boldsymbol{y}_t^{\gamma}=1$。

本文分别设计了在线故障检测、识别和定位3个子任务，以主任务模型与通用解释模型配合的方式实现；其中，故障检测与识别被设计为主任务。首先，针对故障检测任务，利用高维采样数据学习具有单点更新能力的关键变量预测模型$f_t(\cdot)$，则t时刻模型对变量r ($r=1,2,\cdots,R$)的状态估计结果见式(1)。

$$\hat{\boldsymbol{x}}_t^r=f_t^r(\boldsymbol{x}_t) \tag{1}$$

进一步，通过设计状态估计残差的评价函数$g(\cdot)$可得式(2)所示的故障检测结果。

$$\hat{\boldsymbol{y}}_t=g(e_t,\varepsilon_t) \tag{2}$$

其中，e_t通过状态估计误差计算获得，ε_t表示故障报警的阈值。

针对故障识别任务，构建在线学习的故障分类模型h_t，则系统运行状态$\hat{\boldsymbol{y}}_t$通过式(3)判断。

$$\hat{\boldsymbol{y}}_t=h_t(\boldsymbol{x}_t) \tag{3}$$

针对故障定位任务，基于可解释机器学习方法构建面向数据流的特征重要性评价函数$E(\cdot)$，则t时刻故障变量定位的结果形式如式(4)所示。

$$\phi_t=E(L_t,\boldsymbol{x}_t,\boldsymbol{y}_t) \tag{4}$$

其中，L_t表示故障检测或识别模型的损失函数。由式(4)可见，$E(\cdot)$仅基于t时刻主任务模型性能变化筛选故障相关特征，而与主任务模型的网络结构及映射方式无关。

本文所构建的在线故障诊断框架如图1所示，包括基于自适应随机森林(Adaptive Ran-dom Forest，ARF)的故障模式推断模块、特征重要性度量更新模块以及多源报警信息融合模块。该框架首先通过ARF对实时数据流进行回归或分类，并基于ARF的输出判别系统的运行状态，实现故障检测与识别任务。同时，特征重要性度量更新模块基于iPFI在线评估每个传感器特征对于系统状态的全局重要性及其变化率，形成实时的解释信息流，实现故障定位任务。若ARF故障模式推断模块判断系统处于异常或故障状态，特征重要性度量更新模块将联动输出与故障强相关的传感器定位结果。与此同时，多源报警信息融合模块对ARF模块的故障分类判别结果和特征重要性模块的定位信息进行交叉验证与置信度融合，最终输出包含故障类型与故障报警原因的综合诊断结果。

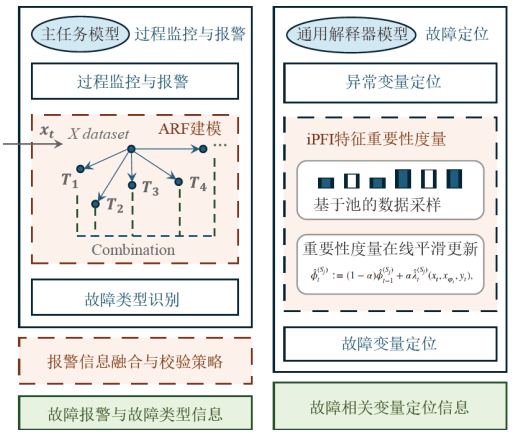


图 1　SA-TCS在线故障诊断框架结构图

Fig. 1　Online fault diagnosis framework for SA-TCS

3 在线故障检测与定位方法建模

3.1 基于ARF的在线故障检测与识别

3.1.1 ARF模型构建

ARF融合了在线学习机制与概念漂移适应策略，是一种能够有效处理演化数据流的集成学习方法。首先，基于Hoeffding Tree构建Q个基学习器，其中每个基学习模型T_i $(i=1,2,\cdots,Q)$利用t时刻采样的传感器数据$(\boldsymbol{x}_t,\boldsymbol{y}_t)$进行迭代训练，训练次数可表示为$k(k\sim\text{Poisson}(\lambda))$。在系统工况切换和状态变化导致数据流发生概念漂移的情况下，集成ARF的每个基学习器T_i通过预警和漂移2个阶段构建变化检测器D_i $(i=1,2,\cdots,Q)$。

在预警阶段，当T_i的预测误差等性能的统计指标发生显著变化并超过预设的预警阈值σ_w时，ARF初始化新的基学习器T_{new}作为背景树，在后续数据流计算中，T_{new}与$T_1\sim T_Q$并行参与训练，但不参与集成投票决策。

在漂移阶段，若T_i累积性能指标变化超过漂移阈值σ_d $(\sigma_d>\sigma_w)$，则认为发生概念漂移，此时用背景树T_{new}替换原有基学习器T_i。

在每个时刻t，集成的第i个分类器$T_{i,t}$在$t+1$时刻的状态$T_{i,t+1}$由检测器状态$S(D_i)$决定，其自适应更新过程如式(5)所示。

$$T_{i,t+1}=\begin{cases}Replace(T_{i,t},T_{i,\text{new}}), & \text{if } S(D_i)=\text{Drift}\\ Initialize(T_{i,t},T_{i,\text{new}}), & \text{if } S(D_i)=\text{Warning}\\ T_i, & \text{otherwise}\end{cases}\tag{5}$$

其中，$Replace(\cdot)$表示对基学习器的替换处理，$Initialize(T_{i,t},T_{i,\text{new}})$表示初始化并开始训练新的背景树。式(5)所示的自适应更新策略使集成模型平滑地适应新数据分布，能够避免因直接重置学习器而导致的模型性能瞬时下降。

3.1.2 在线故障诊断与模型更新

在故障检测阶段，基于正常数据建立单变量预测与在线学习模型，并设计故障检测动态阈值。采用回归模型作为基学习器T_i，基于式(5)构建的ARF基础集成模型f_t^r可对关键变量进行在线状态估计。在任意采样时刻t，变量预测模型f_t^r的输出如式(1)所示。进一步，如式(6)和式(7)所示，通过监测$\hat{\boldsymbol{x}}_t^r$的均方误差是否超过检测阈值，可实现SA-TCS异常及故障检测。f_t^r预测性能的急剧下降指示故障发生。

$$E_t=\frac{1}{L}\sum_{i=t-L+1}^{t}(\boldsymbol{x}_t^r-\hat{\boldsymbol{x}}_t^r)^2\tag{6}$$

$$Fault_Flag_t=\begin{cases}1 & \text{if } E_t>\varepsilon_t\\ 0 & \text{otherwise}\end{cases}\tag{7}$$

其中，L表示异常判断的时间窗长度，动态阈值$\varepsilon_t=\mu_{E,t}+\beta\sigma_{E,t}$，$\mu_{E,t}$和$\sigma_{E,t}$分别表示历史窗口内误差的移动平均值和标准差，$\beta$为可调整参数。

进一步，特别考虑SA-TCS在轨运行过程中已积累的部分故障数据(或模拟数据)，可建立ARF故障识别模型h_t。采用分类模型作为式(5)中的基学习器T_i，在检测故障发生后，仍基于式(5)进一步识别故障的类型。

值得注意的是，在任意时刻t，f_t^r和h_t根据新采样数据$(\boldsymbol{x}_t,\boldsymbol{y}_t)$分别更新为$f_{t+1}^r$和$h_{t+1}$，上述主任务模型的核心功能为实时预测系统的健康状态。在本文所提出的在线故障诊断框架中，定位算法基于模型无关的机器学习可解释性理论实现，因而能够在一致的度量方式下匹配解释任意的增量学习主模型f_t^r和h_t。本文以ARF为例构建了主模型，在后续工作中，将面向数据流特性研究基于自编码或宽度学习的在线学习机制，对主模型结构做出针对性改进。

3.2 基于置换特征重要性的在线故障定位

3.2.1 基于置换特征重要性排序的故障定位

基于PFI算法的模型解释机制，可识别与系统状态关联程度高的关键变量，进而实现故障定位。PFI随机置换特征的实际采样值，并通过观察模型性能下降程度评估该特征的重要性。若某特征对模型的正确预测至关重要，置换该特征将显著增加模型的预测误差。令$(\boldsymbol{x}_1,\boldsymbol{y}_1),\cdots,(\boldsymbol{x}_N,\boldsymbol{y}_N)$表示顺序采样的数据，其中$y_*$表示状态标签或状态估计的真实采样值；$\bar{S}_j$表示除特征$j$以外的所有原始采样值，$S_j$为特征$j$随机置换排序后的数值序列，则基于置换排序的特征$j$的重要度期望如式(8)所示。

$$\bar{\phi}^{(S_j)}:=\frac{1}{N(N-1)}\sum_{n=1}^{N}\sum_{m\neq n}\left\|h\left(\boldsymbol{x}_n^{(\bar{S}_j)},\boldsymbol{x}_m^{(S_j)}\right)-\boldsymbol{y}_n\right\|-\frac{1}{N}\sum_{n=1}^{N}\|h(\boldsymbol{x}_n)-\boldsymbol{y}_n\|\tag{8}$$

其中，N表示样本数量，$h(\cdot)$表示任意主模型函数。由上式可见，PFI期望值本质上衡量了特征置换后模型误差与基于所有特征实例的模型平均误差之间的差值。实际上，PFI的期望$\bar{\phi}^{(s_j)}$可通过设计采样过程φ近似计算，如式(9)所示。

$$\bar{\phi}^{(s_j)}=\frac{N}{N-1}E_{\varphi unif(\mathcal{S}_N)}\left[\hat{\phi}_{\varphi}^{(s_j)}\right]\approx\hat{\phi}^{(s_j)} \tag{9}$$

其中，集合$\mathcal{S}_N$表示样本随机排列的所有可能组合，$E_{\varphi unif(\mathcal{S}_N)}[\cdot]$表示均匀分布采样假设下的期望。

3.2.2　面向数据流的特征重要性在线计算

进一步，将PFI算法推广至在线更新的形式，构建iPFI算法。iPFI与基于在线学习的主模型(回归或分类)并行运行，在每个时刻t针对输入模型的特征j计算并更新其增量特征重要性iPFI_t^j，以实现故障变量定位。若系统中某部件发生故障，其相应的传感器量测数据与系统健康状态的动态关系将发生变化，iPFI的趋势信息也随之显著变化。如式(10)所示，iPFI将全局PFI分解到每个时刻计算，并进行时间加权平均。

定义 单样本重要性估计: 在t时刻，对于特征j，其单样本重要性估计$\hat{\lambda}_t^{(s_j)}$定义见式(10)。

$$\hat{\lambda}_t^{(s_j)}(\boldsymbol{x}_t,\boldsymbol{x}_{\varphi_t},\boldsymbol{y}_t):=\frac{\left\|h_t\left(\boldsymbol{x}_t^{(\bar{s}_j)},\boldsymbol{x}_{\varphi_t}^{(s_j)}\right)-\boldsymbol{y}_t\right\|}{\|h_t(\boldsymbol{x}_t)-\boldsymbol{y}_t\|} \tag{10}$$

其中，φ_t表示从$\{0,1,\cdots,t-1\}$中选择历史样本索引的采样策略，$\boldsymbol{x}_t^{(\bar{s}_j)}$表示$\boldsymbol{x}_t$中除特征$j$以外的所有特征值，$\boldsymbol{x}_{\varphi_t}^{(s_j)}$表示从历史样本$\boldsymbol{x}_{\varphi_t}$中采样的特征$j$的数值，$\|\cdot\|$表示损失函数，如分类任务中的0~1损失或回归任务中的均方误差。

基于式(10)，采用随机的历史数值替换当前样本特征j的真实采样值，并衡量模型预测误差变化。若误差显著增大，表明特征j对于当前主任务模型影响显著。进一步，给出递归形式的iPFI单点更新策略，见式(11)。

$$\hat{\phi}_t^{(s_j)}:=(1-\alpha)\hat{\phi}_{t-1}^{(s_j)}+\alpha\hat{\lambda}_t^{(s_j)}(\boldsymbol{x}_t,\boldsymbol{x}_{\varphi_t},\boldsymbol{y}_t) \tag{11}$$

其中，$t>t_0$且$\hat{\phi}_{t_0-1}^{(s_j)}:=0$，$\alpha(\alpha\in(0,1))$为平滑因子，控制算法对历史数据的记忆长度和对新变化的反应速度，其选择可通过滑动窗口L确定(即$a=1/(L+1)$)。

进一步，为减小特征置换过程计算代价，基于几何采样(Geometric Sampling，GS)设计增量采样策略，以选择历史样本$\boldsymbol{x}_{\varphi_t}$。在GS策略下，通过更新策略保证新样本具有更高的概率替换旧样本，即越接近当前时刻的样本被采集的概率越高。对于SA-TCS，工况切换和故障发生均可导致数据分布改变，GS采样策略使解释器更关注系统当前的行为模式，快速遗忘与当前状态相关性弱的历史信息。在此基础上，iPFI算法的在线更新流程如表1所示。通过持续执行iPFI更新步骤，能够得到并行的特征重要性时间序列，即解释信息流。

表 1　iPFI算法的计算过程

Table 1　Computational process of the iPFI algorithm

算法: iPFI解释器在任意时刻对各输入特征的解释过程	
输入：α, GS采样策略$\hat{\phi}_t^{(s_j)}$, 当前时刻模型h_t, 当前观测样本$(\boldsymbol{x}_t,\boldsymbol{y}_t)$ 输出：	
1	Begin Explain One$(h_t,\boldsymbol{x}_t,\boldsymbol{y}_t,j)$
2	执行采样策略：$x_{\varphi_t}\leftarrow$ Sample(φ_t)
3	根据公式(9)计算单样本重要性$\hat{\lambda}_t^{(s_j)}$
4	根据公式(10)平滑更新获得$\hat{\phi}_t^{(s_j)}$
5	更新采样器：$\varphi_{t+1}\leftarrow$ UpdateSampler$(\varphi_t,\boldsymbol{x}_t)$
6	End Explain One

故障检测与识别模型输出故障报警信息后，基于特征重要性准则评估故障变量范围。为提高故障诊断结果的稳定性，采用式(12)所示的多源报警信息验证与融合策略。

$$F_{j,t}=\frac{1}{W}\sum_{k=t-W+1}^{t}\hat{\phi}_k^{(s_j)} \tag{12}$$

其中，$\hat{\phi}_k^{(s_j)}$表示变量j在k时刻的增量特征重要性，W为时间窗大小。基于时间窗口内各特征重要性的累积值，可识别导致系统偏离正常运行模式的关键监测变量。

4　SA-TCS过程监控与故障定位实验案例分析

4.1　实验对象与实验设计

为验证本文所提出的在线故障诊断框架的有效性，针对某真实的SA-TCS展开实验验证，其结构如图2所示。采用该SA-TCS在轨遥测数据和模拟故障

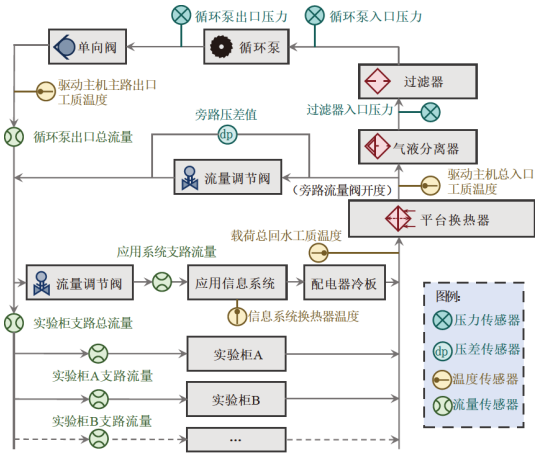


图 2 某SA-TCS结构示意图

Fig. 2 Schematic diagram of SA-TCS

数据开展故障检测、识别与定位子任务的对比实验。该SA-TCS的核心功能是利用循环流动的冷却工质，高效、稳定地收集站内各电子设备和实验载荷产生的废热，并进行冷却作用，以维持航天器有效载荷的各部件在安全的温度范围内工作。SA-TCS具备典型的非线性、强耦合的多工况运行特性。对地面接收的SA-TCS原始遥测数据进行垂直偏移的标准化处理后的数据趋势如图3所示，其中变量名称可与传感器采样的系统部件位置相对应，本文针对原始采样数据可视化均采用数据幅值垂直偏移标准化处理。可见，SA-TCS正常运行状态下的传感器采样数据模式多样，异常或故障事件可导致多维变量数据趋势发生变化。

实验所采用原始遥测数据的采样频率为2 帧/s。

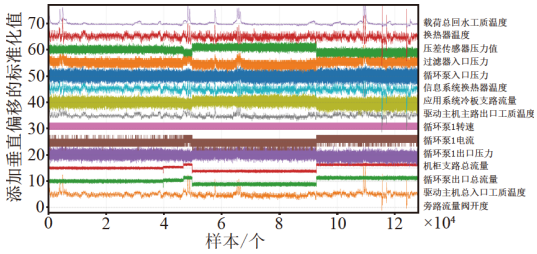


图 3 传感器原始采样数据可视化

Fig. 3 Raw sensor sampling data

后续实验考虑的典型异常(故障)场景如表2所示。此外，基于图3的真实遥测数据，还选取其平稳工况数据构建了模拟故障数据集，所模拟的故障形式及相应变量变化趋势如表3所示。其中，故障变量的偏移程度选取为20%。在后续实验中，首先采用降采样后的真实数据进行了状态估计与故障定位任务的对比实验；进一步，采用模拟故障数据进行了故障检测、识别与定位任务的对比实验。在数据预处理阶段，对各维变量进行了如式(13)的归一化处理。

$$y_{\text{nom}} = \frac{\eta(y - y_{\min})}{y_{\max} - y_{\min}} \tag{13}$$

其中，$y_{\min}$和$y_{\max}$分别表示变量历史数据的最小值和最大值。

4.2 多维监控变量的状态估计与在线异常定位实验

本节实验将关键变量状态估计作为主任务，即采用ARF回归模型对目标变量进行一步预测，以变量重构结果的均方误差作为损失函数；与此同时，解释器依据回归模型损失函数的变化实时评估特征重要度；在异常报警时，解释器依据特征重要度排序输

表 2 实验场景说明

Table 2 Overview of experiment scenarios

实验场景	预测变量	异常发生事件(样本点)	相关变量
实验E1	循环泵出口总流量	2228	ξ_1={旁路压差值, 过滤器入口压力,应用系统支路流量,驱动主机主路出口工质温度, 循环泵电流, 循环泵出口压力}
实验E2-a	载荷总回水工质温度	4980	ξ_2={平台换热器温度, 信息系统换热器温度, 驱动主机主路出口工质温度, 驱动主机总入口工质温度}
实验E2-b	循环泵出口总流量	6800	ξ_3={实验柜支路总流量}
实验E3	载荷总回水工质温度	{7500, 15 100}	ξ_4={平台换热器温度, 过滤器入口压力, 信息系统换热器温度, 驱动主机主路出口工质温度, 旁路流量阀开度}

表 3 模拟故障说明

Table 3 Summary of fault simulation conditions

故障编号	故障类型	故障描述(影响变量)
故障1	SA-TCS驱动组件循环泵转速过高	循环泵转速、循环泵出口压力、应用系统支路流量、循环泵出口总流量、实验柜支路总流量、循环泵电流升高
故障2	SA-TCS驱动组件循环泵泄漏	循环泵出口压力、循环泵出口总流量、应用系统支路流量、实验柜支路总流量降低
故障3	SA-TCS旁路泄漏	循环泵出口压力、循环泵出口总流量、应用系统支路流量、实验柜支路总流量、旁路压差值降低
故障4	SA-TCS驱动组件循环泵转速过低	循环泵电流、载荷总回水工质温度、驱动主机总入口工质温度、驱动主机主路出口工质温度升高；循环泵转速、循环泵出口压力、应用系统支路流量、循环泵出口总流量、实验柜支路总流量降低

出异常变量定位结果。

如表2中的实验E1所示，对原始数据进行1%的降采样处理后，进行在线状态监控，监测数据区间如图4所示。值得注意的是，主任务模型从第一个采样点开始顺序学习，获得每个采样点下一时刻的估计值后，计算损失函数值，并采用当前时刻采样数据在线更新模型参数；在后续实验中，主任务模型均采用上述单点更新的方式进行在线学习。

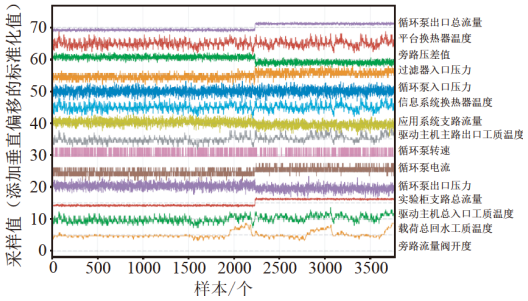


图 4　实验E1监测区间内原始数据可视化

Fig. 4　Raw data obtained within the monitoring interval of experiment E1

在实验E1中，主任务基于ARF对循环泵出口总流量进行状态估计与异常检测，并采用iPFI解释器在异常报警时刻进行实时定位。表2中实验E1记录了异常实际发生时间t为第2228个样本点。根据专家经验与真实数据标注，该异常事件需要关注的变量集合为表2中的ξ_1。图5给出了实验E1场景下的过程监控和故障定位结果。进一步，实验将本文方法与其他模型无关的解释方法进行对比。在相同的主预测任务设定下，采用Shapley累积全局重要性（Shapley Additive Global importancE，SAGE）算法对该预测任务进行增量式特征重要度计算。SAGE通过衡量特征在所有可能特征子集中的平均边际贡献，来分配其对于模型预测的全局重要性[18]，结果见图6。对比图5和图6可知，在异常发生后，主任务检测器的重构误差可立即触发报警，基于本文方法计算的异常事件相关变量的实时特征重要度快速发生阶跃变化，且其定位结果与基于专家经验变量集合ξ_1相比偏差较小；而基于SAGE的特征重要性衡量方法导致各变量重要性计算结果均产生显著波动，难以区分异常的主要影响变量。由实验结果对比可知，基于iPFI方法的异常定位引发干扰变量较少，能够有效解释异常事件的发生原因，为地面专家提供更为可信的异常或故障参考信息。

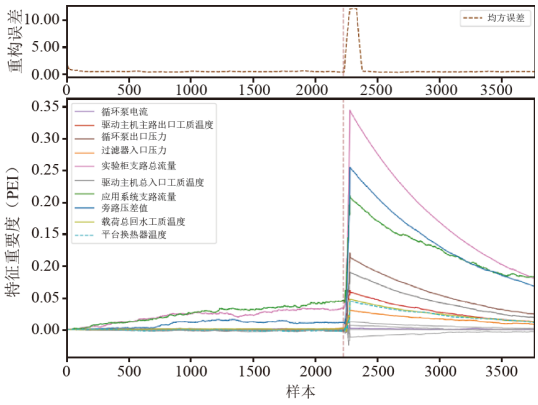


图 5　基于iPFI的实时异常定位结果（实验E1）

Fig. 5　Real-time fault localization results obtained using iPFI for experiment E1

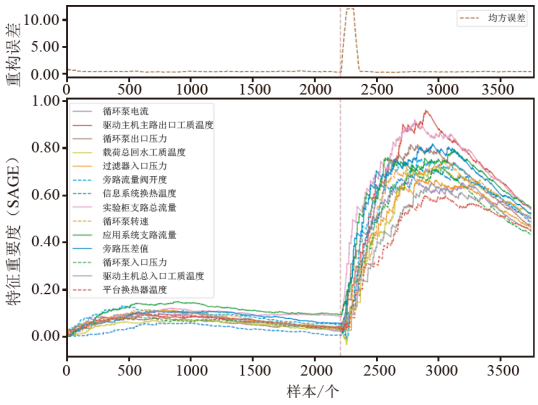


图 6　对比基于SAGE的异常定位结果（实验E1）

Fig. 6　Comparison of SAGE-based fault localization results for experiment E1

进一步，在实验E2-a中对载荷总回水工质温度变量进行状态监控。为了更明显地展示变量的变化趋势，对原始数据以10%降采样处理，数据趋势如图7所示。参考表2中实验E2-a变量集合ξ_2，SA-TCS中某位置工质温度的异常波动随着工质流动对相邻监控节点温度均产生一定程度影响。

图8和图9给出了本文方法与SAGE解释器异常

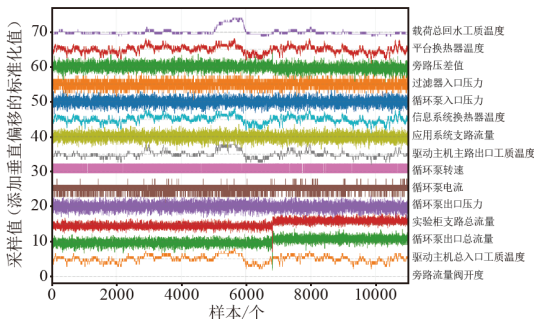


图 7　实验E2监测区间内原始数据可视化

Fig. 7　Raw data obtained within the monitoring interval of experiment E2

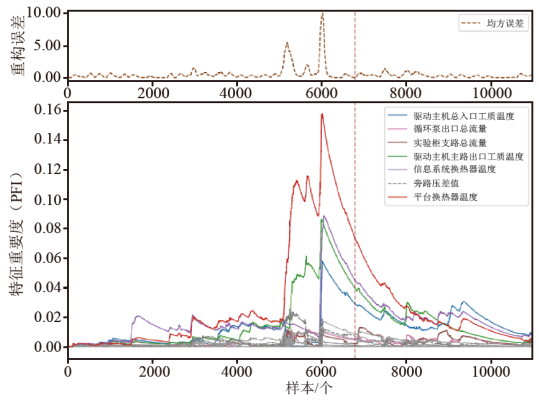


图 8 基于iPFI的实时异常定位结果(实验E2-a)

Fig. 8 Real-time fault localization results obtained using iPFI for experiment E2-a

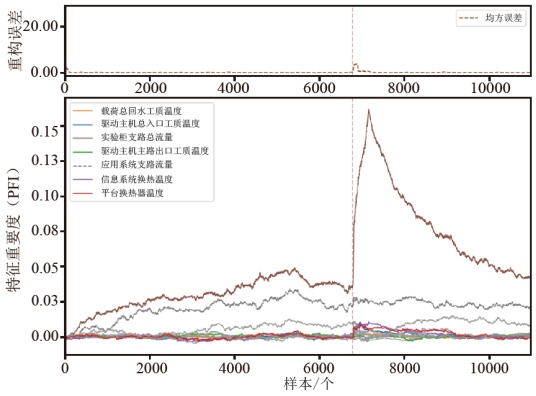


图 10 基于iPFI的实时异常定位结果(实验E2-b)

Fig. 10 Real-time fault localization results obtained using iPFI for experiment E2-b

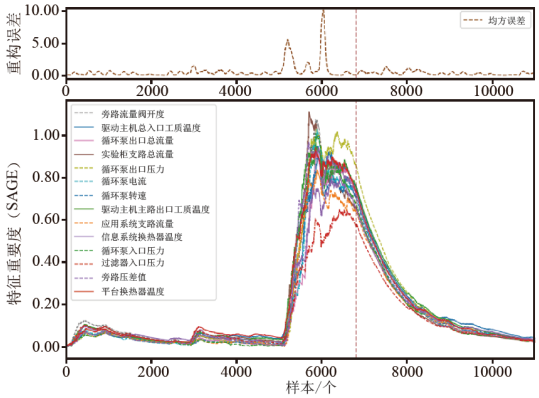


图 9 对比基于SAGE的异常定位结果(实验E2-a)

Fig. 9 Comparison of SAGE-based fault localization results for experiment E2-a

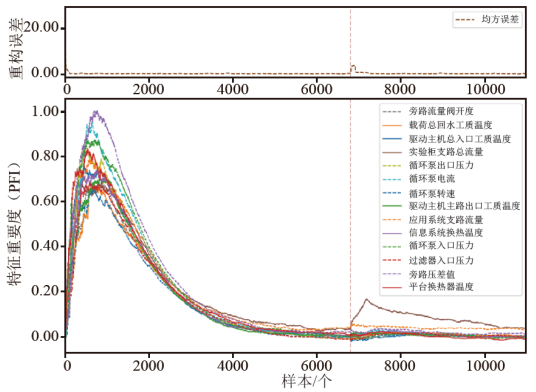


图 11 对比基于SAGE的异常定位结果(实验E2-b)

Fig. 11 Comparison of SAGE-based fault localization results for experiment E2-b

变量定位结果的对比。由实验结果可见,载荷总回水工质温度变量发生异常波动后,采用本文方法能够准确、快速区分与该异常事件相关的温度变量;而SAGE方法无法区分主要影响变量,因而无法实现有效定位。

在此基础上,实验E2-b将循环泵出口总流量作为预测变量,参考图7可见,该预测变量在第6800个采样点发生异常波动,其主要影响变量为实验柜支路总流量。基于iPFI和SAGE方法分别获得异常检测和定位结果如图10和图11所示。图10表示采用本文iPFI方法计算特征重要性,在模型训练更新阶段,所有特征重要性趋势均较为平稳;异常发生后,实验柜支路总流量的特征重要性立即发生显著变化。SAGE算法在训练未收敛阶段即明显受到损失函数变化影响,可能为地面监控人员带来干扰信息;在后续监控过程中,SAGE仅在一定程度上反映异常相关变量的联动趋势,各变量重要性与正常状态下区分度不足。

实验E3对关键变量载荷总回水工质温度进行了预测。由图12所示的原始数据趋势可见,载荷总回水工质温度变量的异常波动可能受到平台换热器温度、信息系统换热器等多节点或部件温度影响,也可能受到流量调整和压力值波动的影响。

如图13和图14所示,采用iPFI方法能够及时区分

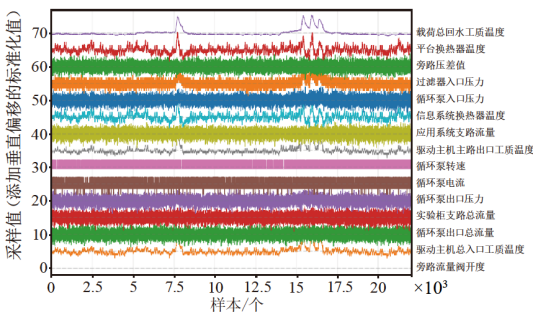


图 12 实验E3监测区间内原始数据可视化

Fig. 12 Raw data obtained within the monitoring interval of experiment E3

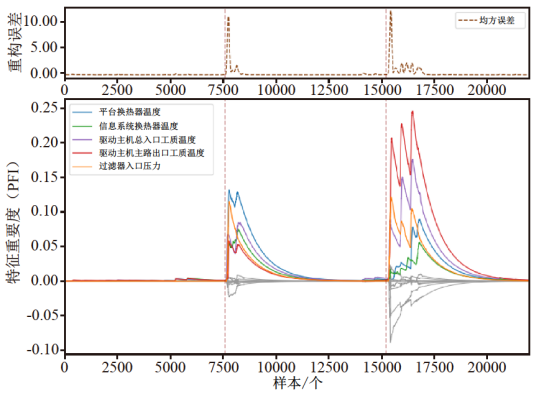


图 13　基于iPFI的实时异常定位结果(实验E3)

Fig. 13　Real-time fault localization results obtained using iPFI for experiment E3

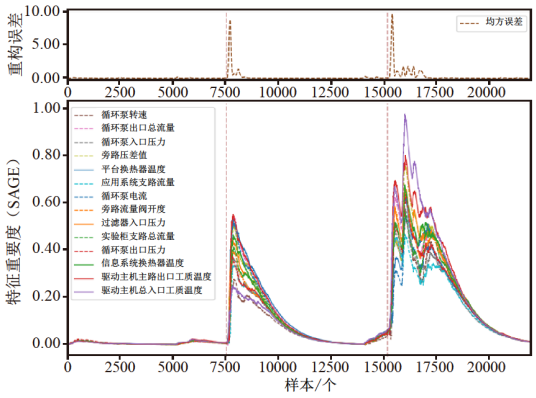


图 14　对比基于SAGE的异常定位结果(实验E3)

Fig. 14　Comparison of SAGE-based fault localization results for experiment E3

所有与异常事件相关和无关的变量集合，为异常溯源提供参考信息；而SAGE方法使各变量对主任务损失函数变化均敏感，未反映异常信息。

在此基础上，实验E4利用了模拟故障模式数据进行故障检测与定位实验。本节实验和后续实验所考虑的故障模式及其影响变量如表3所示。实验E4考虑了正常、故障1和故障3三种运行模式，每种运行模式持续时长为8000个采样点，数据趋势如图15所示。实验针对循环泵出口总流量进行基于单点在线更新的变量状态监控与故障检测，并完成实时变量定位。对比图16和图17实验结果，iPFI方法能够在检测故障发生之后准确、及时地定位与异常变量相关或耦合的变量集合，其结果明显优于基于SAGE方法的故障变量定位结果。

4.3　多维监控变量的在线故障识别与定位实验

本节实验主要将主任务设计为系统故障状态识别，采用ARF分类器模型对高维变量的采样数据进行故障模式分类；解释器依据分类模型的交叉熵损

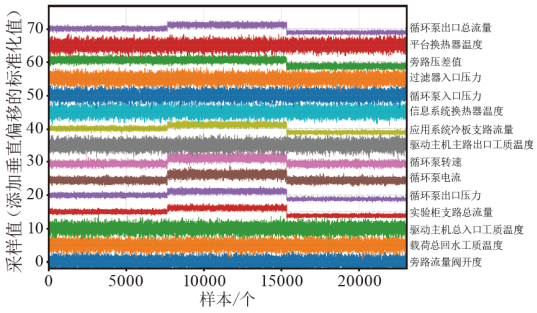


图 15　实验E4模拟故障数据可视化

Fig. 15　Simulated data obtained under fault diagnosis modes of Experiment E4

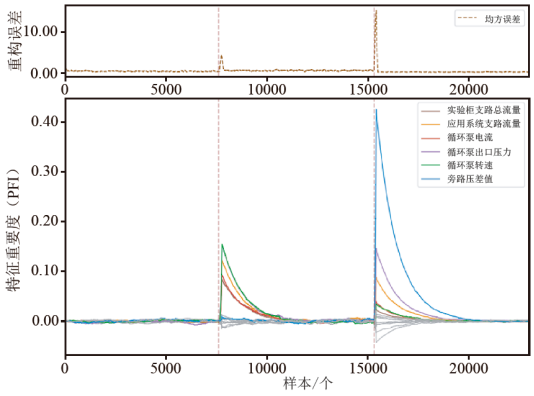


图 16　基于iPFI的实时异常定位结果(实验E4)

Fig. 16　Real-time fault localization results obtained using iPFI for experiment E4

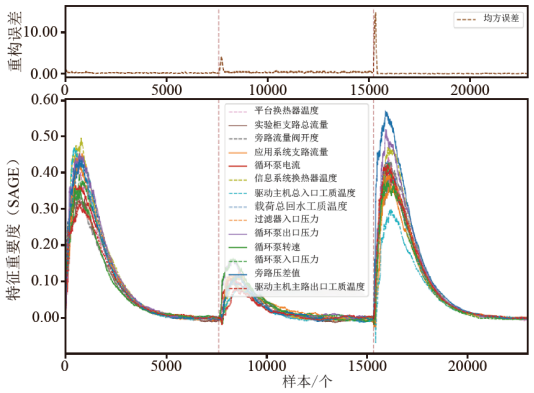


图 17　对比基于SAGE的异常定位结果(实验E4)

Fig. 17　Comparison of SAGE-based fault localization results for experiment E4

失函数变化度量特征重要度，实现故障相关变量定位。

4.3.1　故障检测与定位性能对比

首先在实验场景F1中考虑仅区分正常和故障2种工作模式的故障检测与定位任务。如图18所示，F1实验模拟了系统从正常运行到故障状态的场景，其中故障数据采用与正常状态数据集容量相同的故障1类型的数据。本文在分类任务中均假设故障模

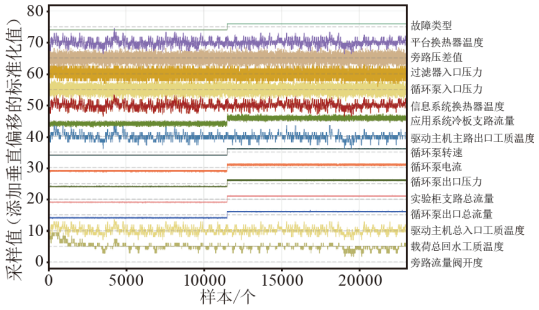


图 18 实验F1模拟故障数据可视化

Fig. 18 Simulated data obtained under fault diagnosis modes of experiment F1

式随机出现，对不同运行模式的数据集进行了随机排序处理，因此基于在线学习的分类器能够有效学习故障模式数据。图19和图20分别给出了基于iPFI方法和SAGE方法的故障定位结果，其子图分别展示了各运行状态下的相关变量。值得注意的是，由于解释器重点关注引起正常与故障模式数据分布差异的特征变量，正常和故障1模式下的重要特征趋势一致，反映了与故障1显著相关的特征集合。对比2种解释方法可见，SAGE方法的稳定性欠佳，iPFI方法则能够给出更为平滑的故障特征估计结果。在后续研究中，将考虑进一步以隶属度的方式表征各变量与故障模式的相关程度，以便为地面专家提供故障影响范围的量化结果。

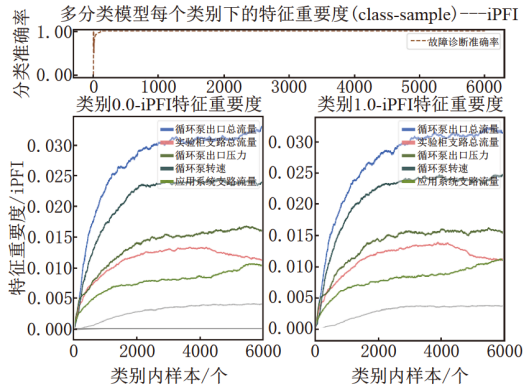


图 19 基于iPFI的实时异常定位结果(实验F1)

Fig. 19 Real-time fault localization results obtained using iPFI for experiment F1

4.3.2 故障识别与定位性能对比

本节实验对比iPFI对多分类任务的解释能力。在实验场景F2~F4中，分类模型对SA-TCS的不同运行状态进行识别，并在故障模式下分别识别故障相关变量。如图21所示，实验场景F2通过数据可视化依次给出了正常运行、故障1和故障2三种状态下的

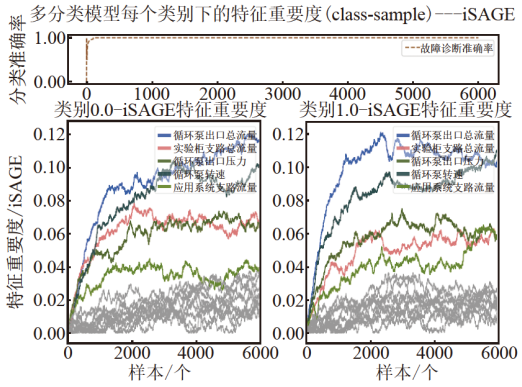


图 20 对比基于SAGE的异常定位结果(实验F1)

Fig. 20 Comparison of SAGE-based fault localization results for experiment F1

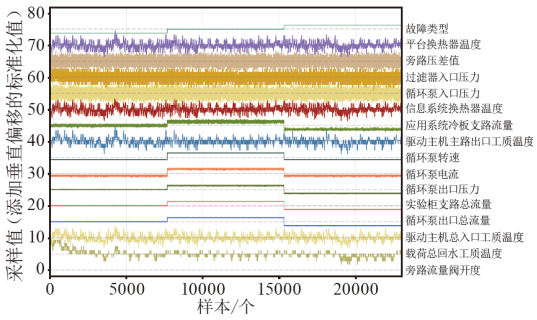


图 21 实验F2模拟故障数据可视化

Fig. 21 Simulated data obtained under fault diagnosis modes of experiment F2

数据趋势。在场景F2中，基于iPFI计算获得的故障1和故障2的相关变量集合如图22所示，其结果与表3中经验变量相符，且具备典型区分度；图23所显示的基于SAGE的计算结果稳定性欠佳，且故障相关变量信息被无关变量掩盖。

图24依次给出了SA-TCS正常运行状态和故障1~故障4状态下的样本数据趋势。在实验场景F3中，故障识别模型对正常状态和故障1~故障3状态的随机数据进行在线学习与分类，基于iPFI和SAGE方法的故障变量识别结果分别如图25和图26所示；在实验场景F4中，故障识别模型对正常运行状态和所有故障状态进行识别，对比方法的实验结果如图27和图28所示。由各类别故障相关变量定位结果对比可见，iPFI方法能够获得更为稳定、精准的故障变量定位结果。

5 结论

本文针对SA-TCS在多工况、长周期运行过程中面临的动态故障诊断挑战，基于ARF与iPFI实现了在

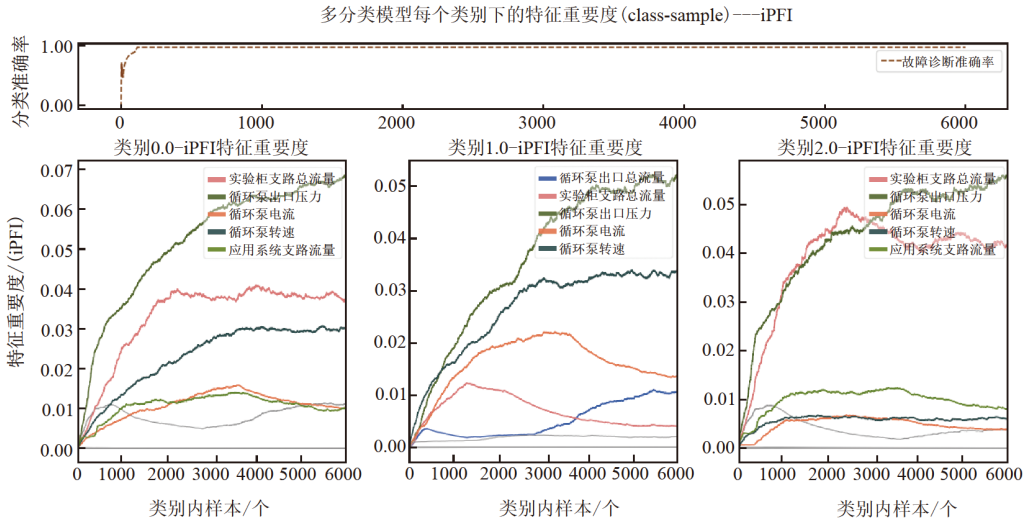


图 22　基于iPFI的实时异常定位结果(实验F2)

Fig. 22　Real-time fault localization results obtained using iPFI for experiment F2

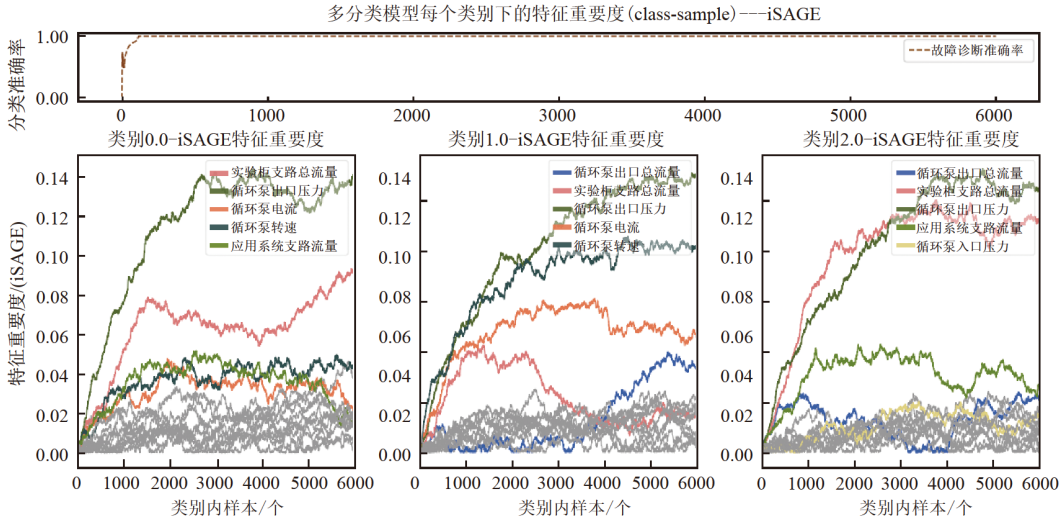


图 23　对比基于SAGE的异常定位结果(实验F2)

Fig. 23　Comparison of SAGE-based fault localization results for experiment F2

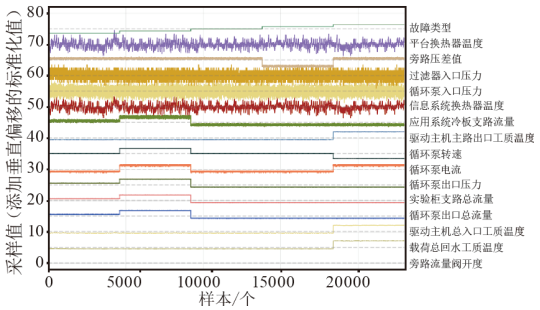


图 24　实验F3模拟故障数据可视化

Fig. 24　Simulated data obtained under fault diagnosis modes of experiment F3

线故障检测与定位,结论如下:

1)提出了一种将在线机器学习模型与在线XAI方法相结合的故障诊断新框架,并在该统一框架下实现了可解释的故障检测、识别与定位。

2)将特征重要性从离线的静态指标推广为在线的动态计算形式。在工况切换与部件故障情境下实现了稳定、实时的故障定位。

3)以某SA-TCS为例,利用其真实异常数据和模拟故障数据验证了所提出的故障诊断框架的准确性与实时性。

未来工作可从三方面开展:①多特征关联分析。目前iPFI主要关注单个特征,可扩展到特征组的重要性分析,以诊断更复杂的关联故障。②与其他XAI方法融合,提供从系统宏观状态到单次异常预测的多层次诊断分析。③将所提出方法部署至真实的SA-TCS测试平台上,验证其在物理系统中的性能和鲁棒

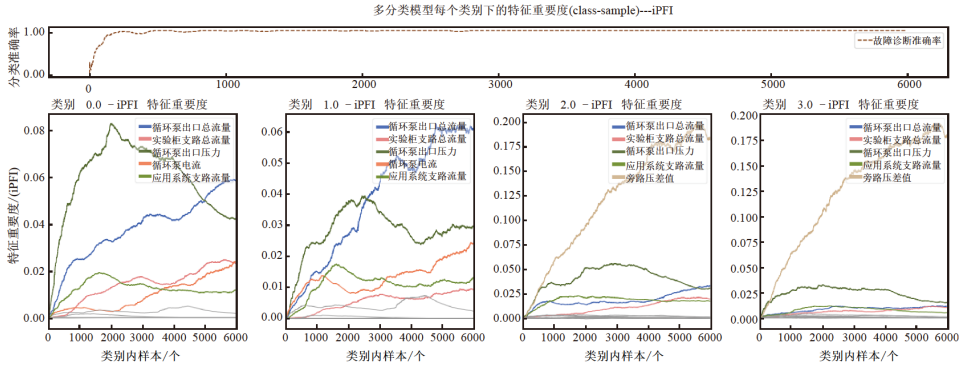


图 25 基于iPFI的实时异常定位结果(实验F3)

Fig. 25 Real-time fault localization results obtained using iPFI for experiment F3

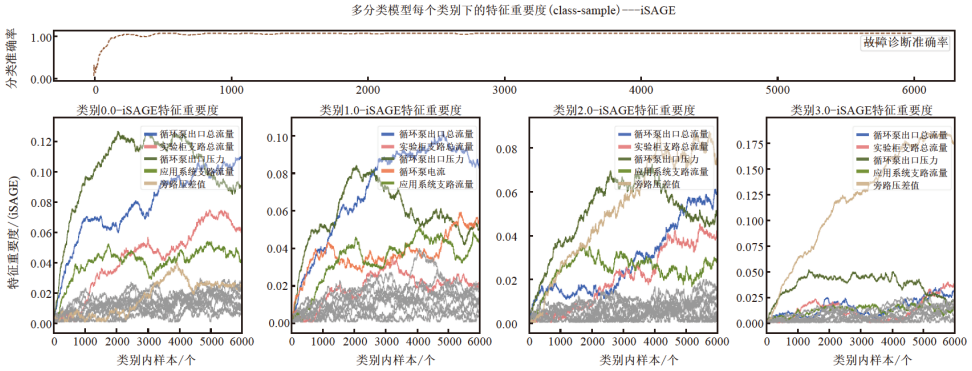


图 26 对比基于SAGE的异常定位结果(实验F3)

Fig. 26 Comparison of SAGE-based fault localization results for experiment F3

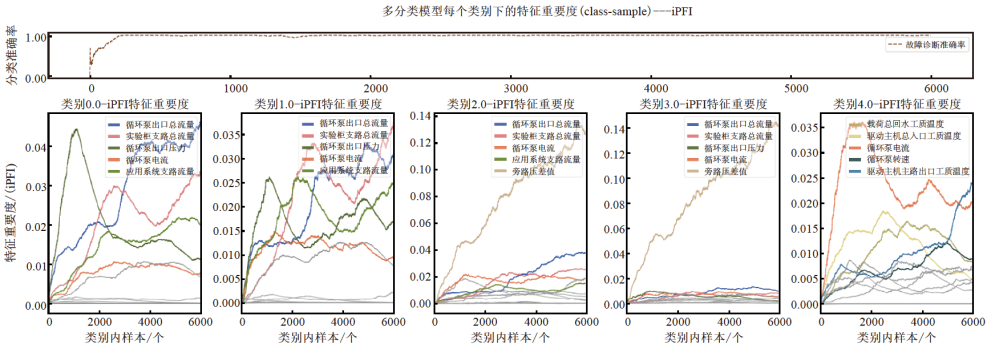


图 27 基于iPFI的实时异常定位结果(实验F4)

Fig. 27 Real-time fault localization results obtained using iPFI for experiment F4

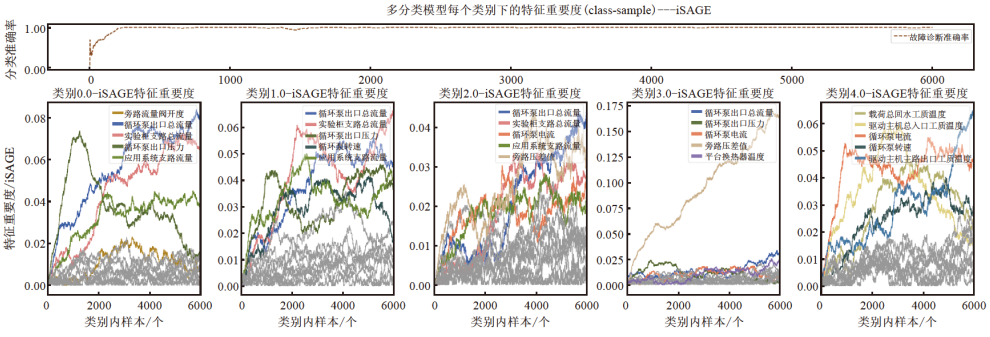


图 28 对比基于SAGE的异常定位结果(实验F4)

Fig. 28 Comparison of SAGE-based fault localization results for experiment F4

性。综上所述，将iPFI这类在线XAI技术引入航天器故障诊断领域，有望推动实现更智能、更透明、更可靠的自主健康管理系统。

参考文献（References）

[1] 张景欣，周东华，陈茂银，等. 数据驱动的多工况过程异常监测方法: 综述与展望[J]. 中国科学: 信息科学，2023，53(11): 2087-2106.
Zhang J X，Zhou D H，Chen M Y，et al. Data-driven anomaly monitoring methods for multimode processes: Overview and perspectives. Sci Sin Inform，2023，53(11): 2087-2106(in Chinese).

[2] Gama J，Žliobaitė I，Bifet A，et al. A survey on concept drift adaptation[J]. ACM Computing Surveys (CSUR)，2014，46(4): 1-37.

[3] Luo X，Wang J，Dong T X，et al. MSFA-Trans: Robust machinery fault diagnosis through multiscale frequency attention transformer in noisy environments[J]. IEEE Transactions on Instrumentation and Measurement，2025，74:1-14.

[4] Hu J，Jia C，Liu H，et al. A survey on state estimation of complex dynamical networks[J]. International Journal of Systems Science，2021，52(16): 3351-3367.

[5] Ju Y，Tian X，Liu H，et al. Fault detection of networked dynamical systems: A survey of trends and techniques[J]. International Journal of Systems Science，2021，52(16): 3390-3409.

[6] Hua X，Huang D，Guo S. Extended state observer based on ADRC of linear system with incipient fault[J]. International Journal of Control Automation and Systems，2019，18(12): 1425-1434.

[7] Pierl D，Vahldiek K，Geißler J，et al. Online model-and data-based leakage localization in district heating networks-Impact of random measurement errors[C]//2020 IEEE International Conference on Systems，Man，and Cybernetics (SMC). Piscataway: IEEE，2020: 2331-2338.

[8] Yin S，Ding S X，Xie X，et al. A review on basic data-driven approaches for industrial process monitoring[J]. IEEE Transactions on Industrial Electronics，2014，61(11): 6418-6428.

[9] Zadkarami M，Shahbazian M，Salahshoor K. Pipeline leakage detection and isolation: An integrated approach of statistical and wavelet feature extraction with multi-layer perceptron neural network (MLPNN)[J]. Journal of Loss Prevention in the Process Industries，2016，43: 479-487.

[10] Kuang J，Xu G，Tao T，et al. Class-imbalance adversarial transfer learning network for cross-domain fault diagnosis with imbalanced data[J]. IEEE Transactions on Instrumentation and Measurement，2021，71: 1-11.

[11] Li X，Duan F，Loukopoulos P，et al. Canonical variable analysis and long short-term memory for fault diagnosis and performance estimation of a centrifugal compressor[J]. Control Engineering Practice，2018，72: 177-191.

[12] Chen K，Hu J，Zhang Y，et al. Fault location in power distribution systems via deep graph convolutional networks[J]. IEEE Journal on Selected Areas in Communications，2019，38(1): 119-131.

[13] Owerko D，Gama F，Ribeiro A. Predicting power outages using graph neural networks[C]//2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP). Piscataway: IEEE，2018: 743-747.

[14] Dong J，Cao K，Peng K. Hierarchical causal graph-based fault root cause diagnosis and propagation path identification for complex industrial process monitoring[J]. IEEE Transactions on Instrumentation and Measurement，2023，72: 1-11.

[15] Hassija V，Chamola V，Mahapatra A，et al. Interpreting black-box models: A review on explainable artificial intelligence[J]. Cognitive Computation，2024，16(1): 45-74.

[16] Kaneko H. Cross-validated permutation feature importance considering correlation between features[J]. Analytical Science Advances，2022，3(9-10): 278-287.

[17] Fumagalli F，Muschalik M，Hüllermeier E，et al. Incremental permutation feature importance (iPFI): Towards online explanations on data streams[J]. Machine Learning，2023，112(12): 4863-4903.

[18] Covert I，Lundberg S，Lee S. Understanding global feature contributions through additive importance measures[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20). Cambridge，MA: MIT Press，2020: 17212-17223.

（责任编辑：高慧）