

基于增量式孤立森林的液体火箭发动机 异常检测方法

张万旋, 薛 薇, 张 楠

(中国航天科技集团有限公司 北京航天动力研究所, 北京 100076)

摘 要: 为解决液体火箭发动机故障标签缺失条件下流数据无监督检测问题, 以及满足不同发动机台次和不同工况的自适应检测需求, 基于增量学习思想, 提出了基于增量式孤立森林的异常检测算法。设计了多工况流数据检测条件下的在线更新策略、异常分数表达式, 并通过更新停止策略避免故障数据对模型的污染。利用多台次试车数据对该模型进行验证, 并与传统方法进行比较, 结果表明, 该算法能够对样本异常程度进行量化评价, 能够有效检测早期缓变故障, 其 F_1 指标较原始孤立森林算法提高了 43%, 检测及时性优于红线算法和自适应阈值算法。

关 键 词: 孤立森林; 自适应检测; 增量学习; 异常检测; 流数据

中图分类号: V434

文献标志码: A

Anomaly detection method of liquid rocket engine based on incremental isolation forest

ZHANG Wanxuan, XUE Wei, ZHANG Nan

(Beijing Aerospace Propulsion Institute,

China Aerospace Science and Technology Corporation, Beijing 100076, China)

Abstract: In order to solve the problem of streaming data unsupervised detection of liquid rocket engine with the absence of fault label, and to enable adaptive detection of various engines and multiple working conditions, an anomaly detection algorithm of liquid rocket engine based on incremental isolation forest was proposed based on incremental learning. The online updating strategy and anomaly score expression for streaming data under various working condition were designed. And the update stop strategy was applied to avoid the pollution of fault on the model. Using a number of test data for analysis and comparison with traditional methods, the result showed that the algorithm can evaluate the anomaly degree of sample quantitatively, and detect the degree of fault effectively. The algorithm had 43% improvement on F_1 score over original isolation forest algorithm. And its detection timeliness was better than the red line algorithm and adaptive threshold algorithm.

Key words: isolation forest; adaptive detection; incremental learning; anomaly detection; streaming data

收稿日期: 2022-03-09

作者简介: 张万旋(1994—), 男, 博士生, 主要研究方向为液体火箭发动机系统设计与故障诊断。

引用格式: 张万旋, 薛薇, 张楠. 基于增量式孤立森林的液体火箭发动机异常检测方法[J]. 航空动力学报, 2022, 37(8): 1674-1682.
ZHANG Wanxuan, XUE Wei, ZHANG Nan. Anomaly detection method of liquid rocket engine based on incremental isolation forest[J]. Journal of Aerospace Power, 2022, 37(8): 1674-1682.

随着航天重大工程的推进,迫切需提高运载火箭动力系统可靠性,液体火箭发动机健康监测与故障诊断技术越来越受到重视,发展灵敏、准确、快速、可靠的故障检测算法,能够在故障早期识别故障并采取措施,避免灾难性后果的产生。发动机通过传感器采集大量、顺序、快速、连续的时序流数据^[1],由于发动机工况变化、不同台次稳态参数偏差、环境干扰等因素的影响,流数据的分布可能产生改变,要求检测算法具有一定自适应能力。故障数据获取的代价是高昂的,发动机数据以正常样本为主,故障样本十分稀缺,这就要求利用大量无标签的数据进行检测,即无监督检测^[2]。

异常是指与正常样本相比满足“少而不同”的样本^[3],由上述分析,液体火箭发动机故障检测可以归结为针对时序流数据在无监督条件下的异常检测问题。

一般来说,异常检测需要对正常数据进行描述,即规定正常样本点的分布范围(距离)或模式(数量或密度),如果某个点不在这个范围内,或者某些点不符合这种分布模式,就会被判定为异常点。常见方法有K-近邻^[4]、聚类^[5]、单分类支持向量机(one class support vector machine)^[6]、基于角度的离群点检测(angle based outlier detection)^[7]、局部异常因子(local outlier factor)^[8]等。20世纪90年代,美国航天飞机主发动机(SSME)的健康监测系统采用K-近邻算法作为故障检测的主要算法之一^[9]。2004年,Iverson^[10]针对SSME的故障检测开发了归纳式健康监测系统(inductive monitoring system),主要采用了基于密度的聚类算法和基于距离的聚类算法。2010年,胡雷^[11]针对液体火箭发动机涡轮泵故障开发了单分类支持向量机检测算法,并通过序贯最小优化算法在线更新模型,提高了算法计算速度。2013年,Schwabacher等^[12]将单分类支持向量机用于SSME数据故障检测。基于距离或密度的异常检测算法有效支撑了发动机故障检测研究,该类方法通常具有以下特点:①需要依据人工经验设置超参数;②需要对数据进行归一化,使各维度数据分布相差不大,从而使模型可以描述特定数据的分布;③距离或密度计算往往需要消耗大量计算资源^[13]。

孤立森林(isolation forest)是Liu等^[14]在2008年提出的一种异常检测算法,该算法利用随机切割隔离所需要的次数来刻画异常程度。与传统异

常检测方法相比,该算法不需要对数据进行归一化处理,并且所需要设置的超参数与数据分布无关(通常越大越好),泛化性能更强;该算法不需要计算距离或者密度指标,具有的 $O(N)$ 时间复杂度, N 为样本数量,并且该算法的基础结构孤立树互相独立,可以通过并行计算的方式提高计算速度^[15]。与自适应阈值算法^[16]等液体火箭发动机故障检测算法相比,该算法能够给出异常程度(通常也是故障程度)的量化评价。

虽然孤立森林问世至今经在众多异常检测算法中的表现名列前茅^[17],但是原始孤立森林算法并不适合流数据。丁智国^[18]提出了一种基于Bagging集成学习的孤立森林流数据异常检测算法,通过对滑窗数据抽样的方式训练新的孤立树,替换异常率偏差较大的孤立树来实现模型在线更新,该算法利用抽样减少了训练规模,但是可能导致信息丢失。李新鹏等^[19]针对电力系统调度过程流数据检测问题,提出改进的孤立森林算法,该算法同样利用滑窗内抽样方法实现流数据检测。针对定向网络攻击流数据检测问题,琚安康^[19]将孤立森林算法与主动学习相结合,通过专家来标记数据集中最有价值的样本,从而减小训练规模,提高检测实时性。贾庆轩^[20]对文献[18]的流数据孤立森林进行改进,引入基于轮盘赌的滑窗缓存改进策略和基于倍增杠杆的异常判别方法。Guha等^[21]提出鲁棒随机割森林(robust random cut forest)算法,该算法可利用单样本对全体孤立树进行更新,但是该算法存在故障数据污染模型等问题。

本文以某型液体火箭发动机为研究对象,基于增量学习思想,提出增量式孤立森林算法:利用新加入的样本在线更新模型,推导适用于增量学习的异常分数表达式,并且通过设置更新停止策略避免故障数据对模型的污染,从而实现液体火箭发动机流数据自适应检测。

1 算法原理

1.1 原始孤立森林算法

孤立森林由多棵孤立树构成,孤立树结构如图1所示。假设 N 是孤立树 T 的一个节点,则它或者是没有父节点的根节点,或者是拥有一个父节点、一个兄弟节点、两个子节点的分支节点,或者是没有子节点的叶子节点。孤立树的高度定义为叶子节点到根节点之间边数的最大值。

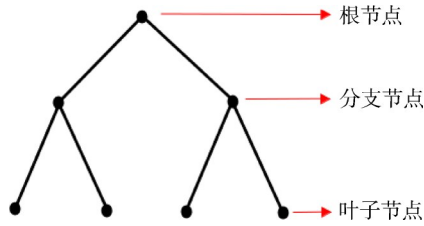


图1 孤立树示意图

Fig.1 Sketch of isolation tree

每一次分割,孤立树随机选择特征 q 和分割值 p ,并满足 $q_{\min} \leq p \leq q_{\max}$, $q_{\min}$ 、 $q_{\max}$ 为特征 q 在数据集 X 中的最小值、最大值。根据样本特征值和分割值的大小关系,将当前样本集合划分到左右两个子节点 X_{left} 、 X_{right} ,具体如算法1。

算法1 生成孤立树 $f_{\text{tree}}(X, e, l)$

输入:数据集 X ,孤立树高度 l ,当前孤立树高度 e

输出:孤立树检测器

1. if $e < l$ and $|X| > l$ then

从 X 的特征集合 Q 随机选择特征 q 和分割值 $p, p \in [q_{\min}, q_{\max}]$

2. X_{left} 中特征 q 的值 $X_{\text{left}} \leq p$

3. X_{right} 中特征 q 的值 $X_{\text{right}} > p$

4. return 枝干节点 N

{ 左子节点 $f_{\text{tree}}(X_{\text{left}}, e+1, l)$,

右子节点 $f_{\text{tree}}(X_{\text{right}}, e+1, l)$,

分割特征 q ,

分割值 p }

5. else

6. return 叶子节点 $N\{X\}$

7. end if

与正常样本点相比,异常样本点更容易被分割到叶子节点,因此可以采用样本点到根节点的深度(即根节点到叶子节点的边数)来表征其异常程度。由于孤立树生成的随机性,单棵孤立树中的样本深度存在波动,需要通过计算样本深度的期望来准确评价其异常程度。具体而言,孤立森林采用多棵孤立树进行独立随机分割计算,并通过样本点 x 在各棵孤立树中深度的平均值来逼近深度的期望,从而计样本点的异常分数 $s(x)$,按下式给出:

$$s(x) = 2 \frac{\frac{1}{t} \sum_{i=1}^t h_i(x)}{c(M)} \quad (1)$$

式中 $h_i(x)$ 为 x 在孤立树 i 中的深度, t 为孤立树数

量, M 为用于生成单棵孤立树的样本点数量, $c(M)$ 起到归一化作用,按下式给出:

$$c(M) = 2 \left[\ln(M-1) + \xi - \frac{M-1}{M} \right] \quad (2)$$

式中 $\xi=0.577\ 215\ 665$ 为欧拉常数。于是,可以给出孤立森林故障检测算法如算法2所示。

算法2 生成孤立森林 $f_{\text{forest}}(X, t, \psi)$ 并进行检测

输入:数据集 X ,孤立树的数量 t ,随机抽样样本数量 ψ

输出:孤立森林检测器 f_{forest} , X 中样本的异常分数 $s(x)$

1. for $i=1$ to t do

$X' \leftarrow$ 对 X 进行随机抽样,采样规模为 ψ

2. $f_{\text{forest}} = f_{\text{forest}} \cup f_{\text{tree}}(X', 0, l)$

3. end for

4. return f_{forest}

5. for $x \in X$ do

6. 根据式(1)计算 $s(x)$

7. return $s(x)$

8. end for

1.2 增量式孤立森林算法

随着数据的加入或删除,增量式模型不断更新。因此,需要给出在模型不断变化情况下数据点异常程度的评价规则,以及数据点加入/删除情况下的模型更新规则。

设数据集 X 中的样本点 y 在孤立树 T 中的深度为 $f(y, X, T)$,则孤立树 T 的复杂度 $m(T)$ 定义为 T 中所有样本点深度的总和^[21],按下式给出:

$$m(T) = \sum_{y \in X} f(y, X, T) \quad (3)$$

对于 X 中的样本点 z ,若 z 是异常点,则删除 X 后,孤立树 T 的复杂度变化较大。因此,可通过复杂度变化的期望来刻画 z 的异常程度,按下式给出:

$$g(z) = E \left[\sum_{y \in X - \{z\}} f(y, X, T) - \sum_{y \in X - \{z\}} f(y, X - \{z\}, T') \right] \quad (4)$$

式中 $E(\cdot)$ 为求期望运算,一般可通过建立多棵孤立树组成孤立森林模型,再对括号内表达式求均值近似计算期望值。为使不同孤立树高度计算得到的异常程度在同一尺度内,对式(4)进行归一

化,并用求均值替代求期望运算,得到样本点 x 的异常分数表达式:

$$s(z) = \frac{1}{l \cdot t} \left[\sum_{y \in X - \{z\}} f(y, X, T) - \sum_{y \in X - \{z\}} f(y, X - \{z\}, T') \right] \quad (5)$$

增量式孤立森林利用新加入的样本点不断更新模型,需要确定新加入样本点对模型划分维度、分割值的影响。原始孤立森林每次划分选取各特征维度的概率是相等的,在引入无关特征(例如,特征值为常数的特征)情况下会降低检测效率。针对该问题,文献[21]采用概率正比于特征极差的方式来选取特征进行划分,使得每次划分波动性较大的特征被选取的概率较大。文献[22]通过左右子树之间的弗雷歇距离对特征排序,排序靠前的特征优先被选取进行划分。文献[23]采用概率正比于特征信息熵的方式来选取特征进行划分,使每次划分时信息熵较大的特征被选取概率较大,该方法需要先对每个维度的参数进行归一化处理。

由于液体火箭发动机各特征参数的量级存在较大差异,如果采用文献[21]的方法,将导致极差数值较大的特征(例如,泵转速)被选取的概率趋近于1,其他特征不会被选取。因此,本文提出采取概率正比于特征相对极差的方式来选取特征,其中,相对极差为极差与额定水平之比。

新分割值 p 由原数据集 X 和新样本点 z 共同确定。如果根据该分割值能够隔离新样本点 z 和原数据集 X ,则新孤立树 T' 一个子节点以下为原孤立树 T ,另一个子节点包含 z 。如果该分割值不能划分 X 与 z ,则舍弃该分割值,将 z 按照 T 中的分割值从根节点往下插入 T ,得到新的孤立树 T' ,具体如算法3。

算法3 插入样本点 $f_{\text{insert}}(T, z)$

输入:孤立树 $T(X)$, X 为样本集合,待插入的样本点 z

输出:插入样本点后的孤立树 $T'(X \cup \{z\})$

1. if $X = \emptyset$ then

return 只含有 z 的叶子节点 T'

2. else

3. X 具有边界框 $B(X) = [x_1^{\text{low}}, x_1^{\text{high}}] \times [x_2^{\text{low}}, x_2^{\text{high}}] \times \cdots \times [x_d^{\text{low}}, x_d^{\text{high}}]$, d 为特征数量, x_i^{low} 、 x_i^{high} 为 X 中特征 i 的最小值、最大值,令 $\hat{x}_i^{\text{low}} = \min \{z_i, x_i^{\text{low}}\}$, $\hat{x}_i^{\text{high}} = \max \{z_i, x_i^{\text{high}}\}$, $\bar{x}_i$ 为 $X \cup \{z\}$

第 i 个维度特征均值, z_i 为 z 的第 i 个特征,并令 $l_i = (\hat{x}_i^{\text{high}} - \hat{x}_i^{\text{low}}) / \bar{x}_i$

4. 随机选取正实数 $r \sim \text{Uniform}\left(0, \sum_{i=1}^d l_i\right)$

5. 计算分割维度 $c = \arg \min \{j | \sum_{i=1}^j l_i \geq r\}$,

并计算分割值 $p = \hat{x}_c^{\text{low}} + \left(\sum_{i=1}^j l_i \bar{x}_i - r \bar{x}_c\right)$.

6. if $p \notin [x_c^{\text{low}}, x_c^{\text{high}}]$ then

重新生成根节点 N

if $z_c \leq p$ then

将 z 划分到 N 的左子节点, X 划分到右子节点

else

将 z 划分到 N 的左子节点, X 划分到右子节点,右子节点以下为原孤立树 T

else

取原孤立树 T 的根节点 N 分割值 p_0

if $z_c \leq p_0$ then

从根节点 N 左子节点往下按照 T 中的分割值依次插入样本点 z

else

从根节点 N 右子节点往下按照 T 中的分割值依次插入样本点 z

7. return 插入样本点 z 后的孤立树 $T'(X \cup \{z\})$

为了节省运算资源,每棵孤立树可以容纳的样本数取有限值 M 。增量式孤立森林算法按照先进先出的原则来删除样本,具体如算法4。

算法4 删除样本点 $f_{\text{delete}}(T, z)$

输入:孤立树 T ,孤立树中的样本点 z

输出:删除样本点 x 后的孤立树 T'

1. 找到 T 中 z 对应的叶子节点 v ,删除 v

2. 设 u 是 v 的兄弟节点,删除 u 和 v 的父亲节点,并用 u 替代父亲节点

3. 从 u 开始,不断删除父亲节点中的样本点 z ,并更新边界框

4. return 更新后的孤立树 T'

随着样本点的加入与删除,多棵孤立树组成的孤立森林模型能够利用式(4)对新加入样本点 z 的异常程度进行评价。但是新加入的样本点将改变孤立森林的结构,使其在随后的检测过程中认为类似 z 的样本点是“正常的”,导致该算法只能定位异常的起止时刻,而对异常阶段样本点异常程度无法做出评价。在故障阶段,这会导致模型被故障数据污染。因此,本文提出设定一个更

新停止阈值 δ ,当样本 z 的异常分数超过 δ ,则仅对 z 进行检测,而不对模型进行更新。

某型发动机采用氧涡轮燃气路的燃气阀对混合比进行阶跃式调节,其主级工作阶段具有低、中、高3个工况。当工况变化时,如果应用更新停止策略,将导致虚警。因此,本文用工况参数 W 表示工况变化, W 取值-1、0、1,分别对应低、中、高3个工况。每当工况发生变化,设定一个缓冲区,认为缓冲区内的新工况样本点都是“正常的”,从而不适用更新停止策略。为避免对变工况后立即发生故障这一情况造成漏警,缓冲区长度不宜过长。经多次试车变工况数据测试发现,缓冲区长度小于4有虚警风险,缓冲区长度大于4无虚警。因此,缓冲区长度大于4可使检测器学习到新工况正常样本的信息。综合考虑算法可靠性与准确性,取缓冲区长度为10(0.1 s)可满足要求。

在算法起始阶段,应对孤立树进行初始化。采用第一个单样本点加小随机数的方式生成若干样本点来初始化。算法5给出了基于增量式孤立森林的液体火箭发动机检测过程。

算法5 生成增量式孤立森林 $f_{\text{forest, incremental}}(X)$ 并进行检测

输入:样本集合 $X=\{x_1, x_2, \dots, x_n\}$

输出:增量式孤立森林 f_{forest} , X 中样本的异常分数 $s(x)$

1. 对孤立森林 f_{forest} 进行初始化, $f_{\text{forest}} = \{T_1, T_2, \dots, T_t\}$

2. for $i=1$ to n

$\forall T \in f_{\text{forest}}, f_{\text{insert}}(T, x_i)$

3. 根据式(4)计算 $s(x_i)$

4. 设更新停止阈值为 δ ,缓冲区长度为 K ,工况参数 $W=\{w_1, w_2, \dots, w_n\}$

5. if $w_i \neq w_{i-1}$ then

$k=1$

else if $k > K$ then

$k=1, \forall T \in f_{\text{forest}}$

if $s(x_i) > \delta$ then

$f_{\text{delete}}(T, x_i)$

if T 的叶子节点数 $>M$, M 为孤立树样本容量, then

$f_{\text{delete}}(T, x_{i-M})$

else

$k=k+1$

6. return $f_{\text{forest}}, \{s(x_i)\}$

2 试车数据分析

试验数据集来自某型氢氧火箭发动机在地面试车,选取发生代表性故障的4次试车数据进行验证,数据集共包含 1.248×10^5 个样本,如表1所示。

表1 试车数据集
Table 1 Test data set

试验	现象	故障定位
1	93~96 s发生故障后恢复	涡轮泵密封环故障
2	突变型破坏	涡轮泵烧蚀故障
3	突变型破坏	涡轮盘故障
4	275 s后参数缓慢下降	涡轮泵叶片故障

2.1 检测参数选取

为了保护人员、产品和试车台安全,某型发动机设置了紧急关机条件。每条紧急关机条件由参数和红线阈值、联判规则构成,如果单个或几个参数越过相应阈值,则执行自动紧急关机。紧急关机条件经过多次试车不断优化,是工程经验与故障教训的结晶。因此,本文参考紧急关机条件相关参数,并根据工程经验适当添加参数,得到14路检测参数如表2所示。

表2 检测参数

Table 2 Detection parameters

参数	符号	参数	符号
氢泵后压力	p_{ef}	氧泵后压力	p_{eo}
燃气发生器氢喷前压力	p_{gif}	燃气发生器氧喷前压力	p_{gio}
推力室氢喷前压力	p_{cif}	推力室氧喷前压力	p_{cio}
氢泵转速	n_{f}	氧泵转速	n_{o}
氢副气蚀管入口压力	p_{wif}	氧副气蚀管入口压力	p_{wio}
氢涡轮入口压力	p_{itf}	氧涡轮入口压力	p_{ito}
推力室压力	p_{c}	燃气发生器压力	p_{g}

2.2 试车数据故障检测

本文增量式孤立森林算法的参数主要有孤立树高度 l 、孤立树数量 t 以及更新停止阈值 δ 。 l 越大,则孤立森林对正常、异常样本的区分能力越强;由大数定律, t 越大,则对异常评分的计算越精确。 l 和 t 越大模型精度越高,但模型训练越复杂。在多种公开数据集上进行测试表明^[14, 18],原始孤

立森林算法取 $l=256$ 和 $t=40$ 足以兼顾模型精度和计算规模。经多次试验发现,由于发动机数据故障信号较为微弱,需要取较大的孤立树高度。因此,综合考虑检测精度与计算速度,取 $l=1\,024$, $t=40$ 。 δ 的选取需要平衡虚警率与漏警率,当 δ 较大,更多的样本被认为“正常”从而被用于更新模型,存在漏警和污染模型风险;当 δ 较小,更少的数据被用于更新模型,模型自适应性减弱,存在虚警风险。利用多次试车数据进行测试,首先,人工打上标签;然后选取较小的 δ 值进行测试,检验虚警和漏警情况,逐步增加 δ 的值,直到不出现虚警和漏警情况。最后,考虑一定裕度,选取 $\delta=0.06$ 。

采用增量式孤立森林算法对试验1~试验4数据进行检测,4次试车整体的平均单样本检测用时为 1.3 ms 。归一化推力室压力曲线和对应的异常评分曲线如图2~图5所示。

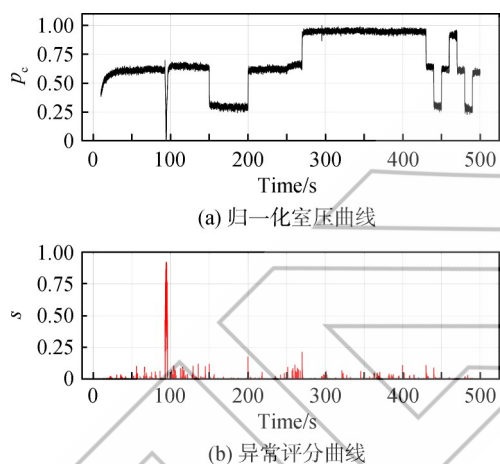


图2 试验1归一化推力室压力和异常评分曲线
Fig.2 Normalized pressure of chamber and anomaly score curves of test 1

2.3 检测结果分析

由图2可知增量式孤立森林模型对4次发生故障的试车数据可进行有效检测。其中,试验1发生故障后,推力室压力等参数仅仅下降了主级水平的4%,但是,故障阶段异常评分(0.92)是正常工作阶段异常评分(经计算其均值为0.00076)的1211倍,因此,该算法对发生参数微弱偏移的故障具有高度敏感性,起到“放大镜”的作用。试验1故障在96s后恢复,异常评分也降低到正常水平。试验2与试验3均为突变破坏故障,经试后分析,性能参数未发现任何故障征兆。针对突变型故障,样本的异常评分垂直上升到接近1,该

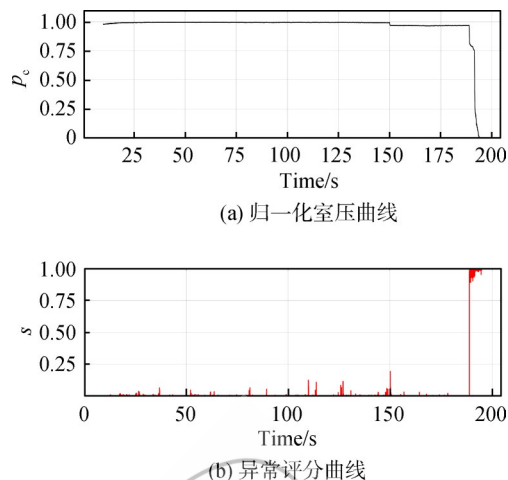


图3 试验2归一化推力室压力和异常评分曲线
Fig.3 Normalized pressure of chamber and anomaly score curves of test 2

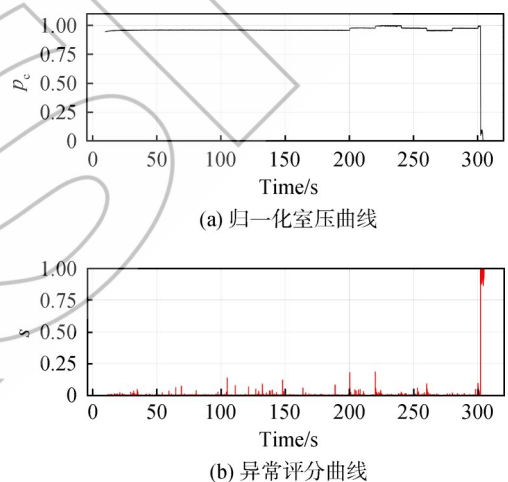


图4 试验3归一化推力室压力和异常评分曲线
Fig.4 Normalized pressure of chamber and anomaly score curves of test 3

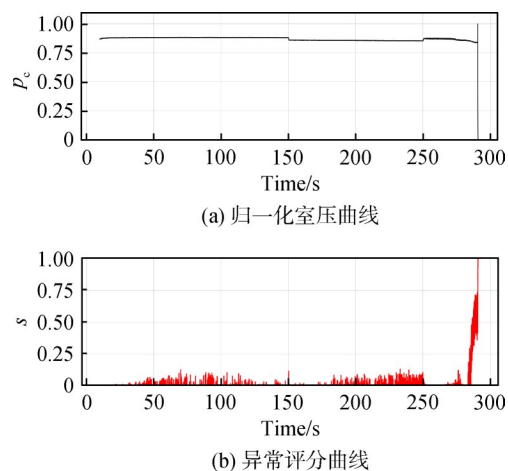


图5 试验4归一化推力室压力和异常评分曲线
Fig.5 Normalized pressure of chamber and anomaly score curves of test 4

算法能够及时进行检测。试验4经试后分析为出现微弱征兆的缓变型故障,并在第290 s发生了严重后果。增量式孤立森林算法给出的异常评分在第285 s就从零附近逐渐上升到1,因此,该算法能够对缓变故障进行有效检测。

由于试验1与试验2为同一台发动机,试验1、试验3与试验4为不同台发动机,因此该算法对于不同台次存在稳态参数差异,同一台次不同工况等情况表现出良好的自适应性。

为量化评估检测结果的准确性,根据试后分析将试车数据打上标签,对于故障发展阶段发动机状态不明确的数据,不纳入评估。原始孤立森林算法将异常判定阈值设置为0.6,但并未给出量化依据^[14]。文献[24]指出,孤立森林算法需通过整体样本分布的先验信息来确定阈值。但是增量式孤立森林算法的检测过程中,进入检测模型的样本总数和分布是变化的,面向在线检测需求也无法获得整体分布先验信息。因此,本文从异常分数定义和工程实际需求出发来确定阈值。根据异常分数的定义,异常分数约等于1减去样本分裂到叶子节点的归一化次数。因此,认为经过20%孤立树高度被分裂到叶子节点的样本为异常样本。由图2,正常工作阶段样本无论处于何种工况,其异常评分几乎都在0.2以下,而故障阶段样本异常评分接近1,因此,本文将异常评分0.2作为异常判定阈值。

根据分类结果和样本标签,有4种可能分类结果:真阳性(T_p)、假阴性(F_n)、真阴性(T_n)、假阳性(F_p),于是可以定义精确率(P)和召回率(R)^[25]:

$$P = \frac{T_p}{T_p + F_p} \quad (6)$$

$$R = \frac{T_p}{T_p + F_n} \quad (7)$$

精确率越大,则虚警越少;召回率越大,则漏警越少,两者是一对矛盾的度量。为了能够综合考虑这两个指标,引入 F_1 指标^[25]:

$$F_1 = 2 \times \frac{PR}{P + R} \quad (8)$$

用 F_1 指标对增量式孤立森林算法检测结果进行评价,并与原始孤立森林算法进行比较。针对3个工况分别建立原始孤立森林模型进行检测,其参数设置与本文算法相同,即 $l=256$, $t=40$,异常判定阈值为0.2。原始孤立森林算法是在离线状态下进行试后检测,使用了全局数据(而不

是流数据),结果见表3,该结果经多次检测取平均值得到。

表3 检测结果评价(F_1 指标)

Table 3 Detection result evaluation (F_1 score)

试验	增量式孤立森林算法	原始孤立森林算法
1	0.847	0.428
2	0.961	0.884
3	0.917	0.819
4	0.496	0.282
综合	0.812	0.564

由表3可知,增量式孤立森林算法能够有效检验4种故障,整体 F_1 指标比原始孤立森林算法提高了43%。由图4,试验1~试验4的虚警主要发生在变工况阶段,多为孤立离群点,可通过设置多点连续检测策略来减少虚警。试验4的 F_1 指标较小是因为检测时间(285 s)晚于标定的故障征兆产生时间(275 s),导致该阶段召回率增大,但是该算法也提供了足够的预警时间。

图6、图7分别给出了试验1和试验4原始孤立森林算法异常评分曲线。原始孤立森林算法检测效果远不及本文算法,原因有四:①对于试验1故障后恢复情况,故障阶段样本较少,可能被抽样所遗漏,造成漏警;②对于试验4缓变故障,单次抽样可能抽取大量故障阶段数据和少量异常数据,不符合异常样本占比较少的前提假设,对缓变

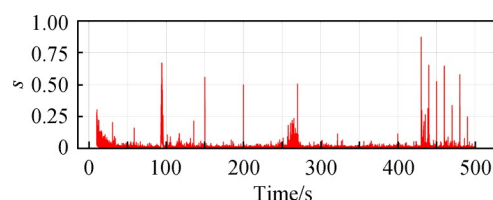


图6 试验1原始孤立森林算法异常评分曲线

Fig.6 Anomaly score curve of test 1 with original isolation forest

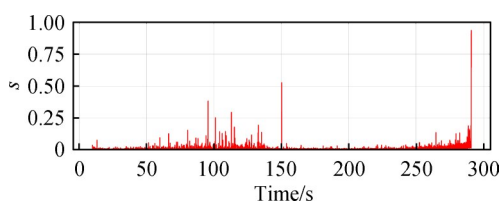


图7 试验4原始孤立森林算法异常评分曲线

Fig.7 Anomaly score curve of test 4 with original isolation forest

故障样本区分能力不强,相比之下,按照时间顺序进行增量学习的方式较容易区分缓变故障;③稳态起始时段参数未平稳时,正常数据被误检;④自适应能力不足带来正常稳态运行阶段虚警。

将增量式孤立森林算法、自适应阈值算法、红线算法进行对比,见表 4,检测结果经多次计算取平均值得到。其中,自适应阈值算法的带宽采用文献[16]的方法来确定。首先,由滑窗内数据的标准差来估计时间序列方差,然后,根据正态分布置信度确定带宽系数,方差估计值与带宽系数相乘即为自适应阈值法带宽,本文选取置信度为 0.999,得到带宽系数为 3.09。红线算法的阈值由该型发动机紧急关机条件确定。

由表 4 可知,针对试验 1 参数下降后恢复和试验 4 缓变故障,本文算法能够在故障早期检测到故障;对于试验 2 和试验 3 突变型故障,本文算法检测及时性略优于自适应阈值算法和红线算法。

表 4 检测时间对比			
Table 4 Comparison of detection result			
试验	增量式孤立森林算法	自适应阈值算法	红线算法
1	92.92	93.31	/
2	188.91	188.92	191.15
3	302.13	302.15	302.19
4	285.26	290.74	290.83

注:“/”表示无法有效检测。

3 结 论

针对液体火箭发动机流数据自适应检测,基于增量学习思想,提出了增量式孤立森林检测算法,经试车数据验证表明该算法检测效果良好。

1) 基于增量式孤立森林算法的故障检测需要先验知识较少,不需要离线训练,对于不同台次发动机、同一台次发动机不同工况、突变故障和缓变故障等情况的检测具有良好自适应性,为实时检测提供了支撑。

2) 与原始孤立森林算法相比,在本文数据集上 F_1 指标提高了 43%,由于按照时间顺序进行增量学习,增量式孤立森林对于缓变故障的区分能力更高,漏警和虚警少,自适应性较好。

3) 增量式孤立森林算法能够对处于发展阶段的故障进行有效检测,检测及时性优于自适应

阈值算法和红线算法。

4) 增量式孤立森林算法能够对异常程度进行量化刻画,故障阶段异常评分是正常阶段的 1 000 倍以上,能够起到“故障放大镜”的作用,为故障程度评价提供了支持。

5) 针对含噪微弱故障,采用特征工程手段来挖掘与故障模式相关的敏感特征,并通过降噪手段进行去噪,增量式孤立森林算法的性能将进一步提高。

参考文献:

[1] 李新鹏,高欣,阎博,等.基于孤立森林算法的电力调度流数据异常检测方法[J].电网技术,2019,43(4):1447-1456.
LI Xinpeng,GAO Xin,YAN Bo,et al.An approach of data anomaly detection in power dispatching streaming data based on isolation forest algorithm[J].Power System Technology,2019,43(4):1447-1456.(in Chinese)

[2] 刘鑫.无监督异常检测方法研究及其应用[D].西安:电子科技大学,2018.
LIU Xin.Research on unsupervised anomaly detection algorithm and application[D].Xi'an:University of Electronic Science and Technology of China,2018.(in Chinese)

[3] 王思齐.智能异常检测及其应用[D].长沙:国防科技大学,2019.
WANG Siqi.Intellegent anomaly detection and its applications[D].Changsha:National University of Defense Technology,2019.(in Chinese)

[4] RAMASWAMY S,RASTOGI R,SHIM K.Efficient algorithms for mining outliers from large data[J].ACM Sigmod Record,2000,29(2):427-438.

[5] ESTER M,KRIEGEL J,SANDER J,et al.A density-based algorithm for discovering clusters in large spatial databases with noise[C]//Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining.Menlo Park,US:AAAI Press,1996:226-231.

[6] SCHÖLKOPF B,WILLIANMSON R,SMOLA A,et al.Support vector method for novelty detection[C]//Advances in Neural Information Processing Systems 12:Proceedings of the 1999 Conference.Cambrtge,US:MIT Press,2000:582-588.

[7] KRIEGEL H P,SCHUBERT M,ZIMEK A.Angle-based outlier detection in high-dimensional data[C]//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.Las Vegas,US:ACM,2008:444-452.

[8] BREUNIG M M,KRIEGEL H P,SANDER N J,et al.LOF:identifying density-based local outliers[C]//Proceedings of the 2000 ACM SIGMOD international conference on Management of data.New York:ACM,2000:1-12.

[9] HAWMAN M W.Health monitoring system for the SSME-program overview[R].AIAA 90-1987,1990.

[10] IVERSON D L.Inductive system health monitoring[C]//

- Proceedings of the International Conference on Artificial Intelligence. Las Vegas, US: ICAI, 2004: 605-611.
- [11] 胡雷. 面向飞行器健康管理的新异常检测方法研究[D]. 长沙: 国防科技大学, 2010.
- HU Lei. Research on novelty detection methods oriented to flight vehicle health management[D]. Changsha: National University of Defense Technology, 2010. (in Chinese)
- [12] SCHWABACHER M, NIKUNJ C O, MATTHEWS B, et al. Unsupervised anomaly detection for liquid-fueled rocket propulsion health monitoring[J]. Journal of Aerospace Computing Information and Communication, 2009, 6: 464-482.
- [13] PEVNY T. Loda: lightweight on-line detector of anomalies[J]. Machine Learning, 2016, 102: 275-304.
- [14] LIU F T, TING K M, ZHOU Z H. Isolation forest[C]// 2008 Eighth IEEE International Conference on Data Mining. Pisa, Italy: IEEE Press, 2008: 413-422.
- [15] 王诚, 狄莹. 孤立森林算法研究及并行化实现[J]. 计算机技术与发展, 2021, 31(6): 13-18.
- WANG Cheng, DI Xuan. Research and parallelization of isolation forest algorithm[J]. Computer Technology and Development, 2021, 31(6): 13-18. (in Chinese)
- [16] 朱恒伟, 王克昌, 陈启智. 液体火箭发动机地面试车故障检测的自适应阈值算法[J]. 推进技术, 2000, 21(1): 1-4.
- ZHU Hengwei, WANG Kechang, CHEN Qizhi. Adaptive thresholds algorithm for fault detection of liquid rocket engine in ground test[J]. Journal of Propulsion Technology, 2000, 21(1): 1-4. (in Chinese)
- [17] LI Z, ZHAO Y, BOTTANICOLA, et al. COPOD: copula-based outlier detection[C]// 2020 IEEE International Conference on Data Mining. Sorrento, Italy: ICDM, 2020: 1118-1123.
- [18] 丁智国. 流数据在线异常检测方法研究[D]. 上海: 上海大学, 2015.
- DING Zhiguo. Research on online anomaly detection for streaming data[D]. Shanghai: Shanghai University, 2015. (in Chinese)
- [19] 琚安康. 基于多源异构数据的定向网络攻击检测关键技术研究[D]. 郑州: 战略支援部队信息工程大学, 2020.
- JU Ankang. Research on key technologies of targeted cyber attacks detection based on multi-source heterogeneous data[D]. Zhengzhou: Information Engineering University, 2020. (in Chinese)
- [20] 贾庆轩. 基于机器学习的电力调度自动化系统业务异常检测方法研究[D]. 北京: 北京邮电大学, 2020.
- JIA Qingxuan. Research on business anomaly detection method of power dispatching automation system based on machine learning[D]. Beijing: Beijing University of Posts and Telecommunications, 2020. (in Chinese)
- [21] GUHA S, MISHRA N, ROY G, et al. Robust random cut forest based anomaly detection on streams[C]// Proceedings of the 33rd International Conference on Machine Learning. New York: Proceedings of Machine Learning Research, 2016: 2712-2721.
- [22] 李国成, 陆俊, 王赞, 等. 基于 Bagging 二次加权集成的孤立森林窃电检测算法[J]. 电力系统自动化, 2022, 46(2): 92-100.
- LI Guocheng, LU Jun, WANG Yun, et al. Isolated-forest electricity theft detection algorithm based on bagging secondary weighted ensemble[J]. Automation of Electric Power Systems, 2022, 46(2): 92-100. (in Chinese)
- [23] 谢炎昆. 基于孤立森林算法的用水异常监测研究[D]. 北京: 北京邮电大学, 2018.
- XIE Yankun. Research on water use anomaly monitoring based on isolated forest algorithm[D]. Beijing: Beijing University of Posts and Telecommunications, 2018. (in Chinese)
- [24] 陈春旭. 基于机器学习的电力调度自动化系统业务异常检测方法研究[D]. 北京: 北京邮电大学, 2020.
- CHEN Chunxu. Research on business anomaly detection method of power dispatching automation systems based on machine learning[D]. Beijing: Beijing University of Posts and Telecommunications, 2018. (in Chinese)
- [25] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016: 403-408.

(编辑: 陈 越)