

基于优化 EIF 的大型望远镜非预期故障诊断*

刘睿强^{1,2,3}, 徐灵哲^{2,3}, 杨世海^{2,3}, 邓壮壮^{1,2,3}, 李运^{2,3}

(1. 中国科学院南京天文光学技术研究所 南京, 210042)

(2. 中国科学院天文光学技术重点实验室 南京, 210042) (3. 中国科学院大学 北京, 100049)

摘要 针对极端环境下的大型望远镜会面临各种无法提前预测的故障和异常情况, 提出一种基于优化扩展孤立森林 (extended isolation forest, 简称 EIF) 的非预期故障诊断方法。首先, 通过特征选择算法 ReliefF 和最大相关最小冗余 (maximum relevance and minimum redundancy, 简称 mRMR) 算法对原始数据集进行特征优选; 其次, 采用蜣螂优化算法 (dung beetle optimizer, 简称 DBO) 为 EIF 算法选择合适参数, 构建最佳故障诊断模型; 最后, 将该模型应用于大型望远镜非预期故障诊断。结果表明: DBO-EIF 算法故障诊断准确率达到 98.7%, 并在 F1 值、Kappa 系数等性能指标上明显优于单类支持向量机 (one-class support vector machine, 简称 OCSVM)、局部离群因子 (local outlier factor, 简称 LOF) 等传统异常检测算法。基于故障诊断及自愈半实物仿真平台证明了 DBO-EIF 算法在非预期故障诊断工程应用中的可行性, 为保证大型望远镜驱动系统在极端环境下的可靠性和稳定性提供了一种有效的解决方案。

关键词 大型望远镜; 驱动系统; 非预期故障; 扩展孤立森林; 特征选择; 蜣螂优化算法

中图分类号 TP391; TH743

引言

为捕捉更深远宇宙信息, 大型望远镜需依赖高精度驱动系统。该系统在复杂结构与极端环境下可能产生偏离既定故障模型的新形态故障, 即非预期故障^[1], 导致性能衰退甚至失效。为保障观测效率与设备安全, 对该类故障的诊断研究势在必行^[2-3]。

为了提高大型望远镜驱动系统非预期故障诊断的精度, 需要选择合适的机器学习分类器。目前, 非预期故障诊断方法可以总结为基于统计、密度、距离、聚类、隔离、集成和子空间等^[4]方法。

从研究方法看, 许多领域都开展了孤立森林 (isolation forest, 简称 iForest)、OCSVM、LOF、支持向量数据描述 (support vector data description, 简称 SVDD) 等异常检测算法的研究。Jiang 等^[5]提出了一种基于孤立森林算法的动力锂电池故障诊断方法, 能够预先检测出渐进式和突发式故障。余翔等^[6]提出了一种基于孤立森林算法的用电数据异常检测方法, 通过实验验证了该算法具有高效准确的优势。Alazzam 等^[7]利用 OCSVM 算法和鸽子启发式优化算法, 结合基于异常检测和入侵检测, 能够有效降低误报率。Kumaravel 等^[8]提出了一种基于 LOF 的鲁棒脑电图数据坏通道检测方法, 能够适应不同的脑

电图数据, 且在 F1 值上优于现有的不良信道检测方法。Wu 等^[9]将 SVDD 算法与流形正则化相结合, 能够有效检测出大规模数据集中的噪声标签, 在离群点检测准确率方面优于其他方法。张万旋等^[10]提出了一种基于增量式 iForest 的异常检测算法, 用于解决液体火箭发动机流数据无监督检测问题, 在检测及时性方面优于传统的红线算法和自适应阈值算法。

从研究对象上看, 目前对大型望远镜驱动系统的非预期故障诊断的研究还较少。李运等^[11]提出了一种基于支持向量机 (support vector machine, 简称 SVM) 状态观测器的天文望远镜新异类故障检测方法, 利用历史运行数据训练 SVM 模型, 通过比较观测输出和实际输出的误差, 判断系统是否处于故障状态, 取得较好的准确率和效率, 但该方法不能微观定位故障, 且样本数据区域覆盖范围有限。蔡杨烁等^[12]提出了一种利用天文定位和轴系定位数据对望远镜状态进行动态监测的方法, 通过分析 2 种数据的误差范围和稳定性, 判断望远镜是否正常工作, 但该方法通过数据残差对比判断望远镜工作状态, 缺乏理论依据和客观标准, 无法适用于非激光卫星和其他空间目标。

* 国家自然科学基金资助项目 (11973065); 天文联合基金重点资助项目 (U1931207); 青年科学基金资助项目 (12303089); 中国科学院海外留学基金资助项目; 国家留学基金管理委员会资助项目 (CSC201904910254)

收稿日期: 2023-08-09; 修回日期: 2023-10-30

为提高大型望远镜驱动系统,笔者提出一种基于EIF的大型望远镜驱动非预期故障诊断方法。通过混合算法ReliefF-mRMR进行特征选择,引入蛭螂优化算法对EIF关键参数进行自适应优化,并在半实物仿真平台上验证了其在诊断精度、稳定性和泛化能力上的优越性。

1 相关工作

1.1 孤立森林算法

孤立森林作为一种无监督异常检测算法^[13],适用于大型望远镜驱动系统的非预期故障诊断。其通过构建多棵孤立树(iTree),依据数据点被隔离的路径长度来度量异常程度,具有效率高、对异常值比例不敏感、无需标签数据即可训练的特点^[14-15]。该算法也存在局限:基于轴平行划分的机制难以契合复杂的数据分布,可能导致对局部异常点的检测失效。图1为二维空间中2种算法异常分数等高线分布,直观揭示了这一差异。

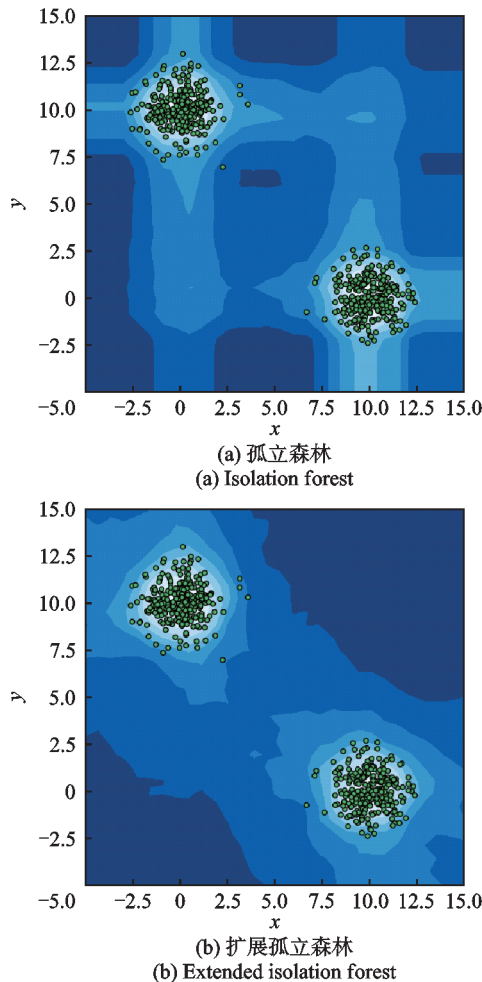


图1 二维空间中2种算法异常分数等高线分布

Fig.1 Contour distribution of abnormal fractions for 2 algorithms in 2 dimensional space

1.2 扩展孤立森林算法

EIF是在iForest的基础上提出的一种改进算法^[16],其核心思想是使用具有随机斜率的超平面来切割数据,这样可以增加切割的多样性和灵活性,提高异常检测的效果。EIF由多棵iTree组成,每棵iTree均为二叉树结构,构建iTree的步骤如下。

1) 初始化参数。对孤立树的数量 t 、子采样大小 p 、树的高度限制 l 、扩展等级 n 等参数进行设置。

2) 对于每一棵树 $i \in [1, t]$ 执行以下操作:①从数据集 X 中随机抽取 p 个样本作为子样本 X_i ;②随机选择一个特征 q 和一个分割值 p ,根据一个随机斜率 ω 和一个随机截距 b ,构造一个孤立超平面。对于 X_i 中的任意一个数据点 x ,若满足 $(x-b)\omega \leq 0$,则将其划分到当前节点的左分支;否则,将其分配到右分支;③重复划分过程,直到满足以下任意一个条件:树达到了高度限制 l ;节点上只有一个样本;节点上样本所有的特征值相同。

3) 利用Ensemble的思想将 t 棵随机生成的孤立树进行集成,构成孤立森林。

在异常检测阶段,给定一个新来的数据 x ,判断其是否为异常点的步骤如下。

1) 对于每一棵孤立树 $i \in [1, t]$,从根节点开始沿切割条件下降至叶子结点,计算数据 x 在孤立树中经过的边数,即路径长度 $h(x_i)$ 。

2) 计算数据 x 在孤立森林中路径长度的期望值 $E(h(x_i))$,这个值反映了 x 被孤立的难易程度, x 越小,表示越容易被孤立,即越有可能是异常点。

3) 数据 x 的异常分数计算式为

$$s(x, p) = 2^{-\frac{E(h(x_i))}{c(p)}} \quad (1)$$

其中: $c(p)$ 为给定样本数 p 时,采用路径长度的平均值来标准化路径长度。

根据式(1)对 x 进行分类。若异常分数越接近1,则表明 x 越异常;若异常分数越接近0,则表明 x 越正常。

$c(p)$ 的计算公式为

$$c(p) = 2H(p-1) - \frac{2(p-1)}{p} \quad (2)$$

其中: $H(p-1)$ 为调和函数, $H(p-1) = \ln(p-1) + 0.5772156649$ (欧拉常数)。

1.3 ReliefF-mRMR混合算法特征选择

笔者引入ReliefF-mRMR混合算法进行特征选择,以提升诊断模型的精度与效率。该方法融合了2种算法的优势:①利用基于距离的ReliefF算法^[17],依据特征对区分同类/异类样本的贡献度计算权重,筛选出高判别性特征,但其未考虑特征间相关性;

②采用基于信息论的 mRMR 算法^[18]从剩余特征中选取与目标类别相关性最大且相互间冗余度最小的特征子集,但其默认各特征贡献均匀。本混合策略先由特征选择算法 ReliefF 剔除低权重特征,再由 mRMR 算法精选特征子集,从而在缓解 EIF 高维数据处理瓶颈的同时,显著降低计算量。

ReliefF-mRMR 混合算法的主要流程如下。

输入:原始数据集 D , 抽样次数 m , 最近邻样本数 k , 目标特征子集大小 s , 特征权重阈值 τ 。

1) 初始化:将所有特征的权重设为 0, 将最优特征子集 S 设为空集。

2) 重复以下操作 m 次:①从 D 中随机抽取一个样本 X ;②从 X 的同类样本集中找到 X 的 k 个最近邻样本 $H_j(j=1, 2, \dots, k)$, 从每一个不同类样本集中找到 k 个最近邻样本 $m_j(c)$;③对于每个特征 A , 更新其权重 $W(A)$, 并剔除权重值小于阈值的特征, 得到特征子集 D_n 和权重矩阵 W_n , 即

$$W(A) = W(A) - \sum_{j=1}^k \frac{\text{diff}(A, x_i, H_j)}{mk} + \sum_{c \neq \text{class}(x_i)} \left\{ \frac{P(c)}{1 - P(\text{class}(x_i))} \frac{\sum_{j=1}^k \text{diff}[A, x_i, m_j(c)]}{mk} \right\} \quad (3)$$

其中: $\text{diff}(A, x_1, x_2)$ 为样本 x_1 和 x_2 在特征 A 上的距离; $P(c)$ 为目标类别 c 的概率; $m_j(c)$ 为目标类别 c 的第 j 个样本。

3) 对特征子集 D_n 进行加权处理, 得到加权后的特征子集 D_m 为

$$D_m = D_n W_n \quad (4)$$

4) 对于每个特征 A , 计算其与类别 c 之间的互信息为

$$I = \frac{1}{|D|} \sum_{x_A \in D} I(x_A; c) \quad (5)$$

其中: $I(x_A; c)$ 为特征 i 和目标类别 c 之间的互信息。

5) 从加权后的特征集中选出与类别 Y 具有最大互信息的一个特征 A^* , 将其加入到 S 中。

6) 根据式(6)从加权后的特征集中选出与类别 Y 具有最大互信息, 且与 S 中已有特征具有最小互信息的一个特征 A^* , 将其加入 S 中, 重复操作 $s-1$ 次, 即

$$\max_{x_i \in D - D_m} \left[I(x_i; c) - \frac{1}{m} \sum_{x_j \in D_m} I(x_i; x_j) \right] \quad (6)$$

输出: 返回最优特征子集 S 。

1.4 蜣螂优化算法

为优化 EIF 模型的关键参数, 提升其对故障数据的诊断精度与效率, 笔者引入 DBO^[19]并与之融

合。DBO 算法将初始种群划分为滚球者、孵卵者、觅食者与偷窃者 4 个子种群, 每个过程都对应一个算子, 用于更新子种群中个体的位置和适应度。通过 4 个相应的算子进行迭代优化, 其核心步骤如下。

1) 滚球过程。模拟蜣螂滚动粪球的全域探索行为, 包含利用天体线索的“滚动”和遇障转向的“跳舞”2 个步骤, 以定位有希望的搜索区域。

滚动步骤: 滚球者分别根据天体线索、光强度变化来确定运动方向和运动速度, 其滚球者的位置更新方式为

$$x_i(t+1) = x_i(t) + \alpha k x_i(t-1) + b \Delta x \quad (7)$$

其中: $x_i(t)$ 为第 i 个滚球者在第 t 次迭代时的位置; k 为偏转系数; Δx 为光强度变化量; α 表示自然因素对运动方向的影响。

Δx 越大, 表示光源越弱, 滚球者越倾向于避开该位置。 Δx 定义为

$$\Delta x = |x_i(t) - X^w| \quad (8)$$

其中: X^w 为当前种群中最差的位置, 即光源最弱的位置。

$\alpha=1$, 表示自然因素不影响原始方向; $\alpha=-1$, 表示偏离原始方向。 α 定义为

$$\alpha = \begin{cases} 1 & (p \geq 0.5) \\ -1 & (p < 0.5) \end{cases} \quad (9)$$

其中: p 为一个服从均匀分布 $U(0, 1)$ 的随机数。

跳舞步骤: 当滚球者遇到障碍物无法前进时, 改变运动方向, 其位置更新方式为

$$x_i(t+1) = x_i(t) + \tan(\theta) |x_i(t) - x_i(t-1)| \quad (10)$$

其中: θ 为偏转角, 服从切线分布 $T(0, \pi)$, 即 $\theta = \arctan(U(-\pi/2, \pi/2))$, 当 $\theta=0, \pi/2$ 或 π 时, 滚球者不会改变位置。

2) 繁殖过程。模拟蜣螂在粪球中产卵的行为, 以增加种群多样性。孵卵者以滚球者找到的优质解区域为中心, 随机生成新个体(产卵)并进行孵化。产卵区域的范围定义为

$$Lb^* = \max(X^*(1-R), Lb) \quad (11)$$

$$Ub^* = \max(X^*(1+R), Ub) \quad (12)$$

其中: Lb^* 和 Ub^* 分别为产卵区域的下界和上界; Lb 和 Ub 分别为搜索空间的下界和上界; X^* 为当前种群中最优的位置, 即产卵区域的中心; $R=1-t/T_{\max}$, 为惯性权重; t 为当前迭代次数; $T_{\max}$ 为最大迭代次数。

新个体的位置定义为

$$B_i(t+1) = X^* + b_1(B_i(t) - Lb^*) + b_2(B_i(t) - Ub^*) \quad (13)$$

其中: B_i 为第 i 个新个体的位置; b_1 和 b_2 为 2 个服从均匀分布 $U(0, 1)$ 的随机向量。

3) 觅食过程。模拟幼蛱螂觅食行为,以提升局部开发能力与收敛速度。觅食者围绕全局最优解进行局部精细搜索。觅食区域的范围定义为

$$Lb^b = \max(X^b(1-R), Lb) \quad (14)$$

$$Ub^b = \max(X^b(1+R), Ub) \quad (15)$$

其中: Lb^b 和 Ub^b 分别为觅食区域的下界和上界; X^b 为全局最优解,即觅食区域的中心; R 为惯性权重,与繁殖过程中的定义相同。

觅食者的位置更新方式为

$$x_i(t+1) = x_i(t) + C_1(x_i(t) - Lb^b) + C_2(x_i(t) - Ub^b) \quad (16)$$

其中: $x_i(t)$ 为第 i 个觅食者在第 t 次迭代时的位置; C_1 为一个服从正态分布 $N(0,1)$ 的随机数; C_2 为一个服从均匀分布 $U(0,1)$ 的随机向量。

4) 偷窃过程。模拟蛱螂抢夺粪球行为,以帮助种群跳出局部最优。偷窃者在全局最优解附近进行随机扰动,其位置更新方式为

$$x_i(t+1) = x^b + Sg[(x_i(t) - X^*) + (x_i(t) - X^b)] \quad (17)$$

其中: $x_i(t)$ 为第 i 个偷窃者在第 t 次迭代时的位置; S 为常数; g 为服从正态分布 $N(0,1)$ 的随机向量。

各子种群在迭代中动态调整,通过上述行为的协同作用实现对EIF参数的高效寻优。

1.5 DBO优化的扩展孤立森林算法

本算法将EIF的4个关键参数(孤立树数量 t 、子采样大小 p 、树的高度限制 l 和扩展等级 n)映射为蛱螂个体的位置信息。针对数据类别不平衡问题,适应度函数采用训练集上五折交叉验证的平均Kappa系数的负值,以避免模型偏向多数类。Kappa系数的计算公式为

$$k = \frac{p_0 - p_e}{1 - p_e} \quad (18)$$

其中: p_0 为总体分类精度,即每类正确分类数与总样本数的比值。

假设每类真实样本数为 $a_1, a_2, \dots, a_c$,每类预测样本数为 $b_1, b_2, \dots, b_c$,则

$$p_e = \frac{a_1 b_1 + a_2 b_2 + \dots + a_c b_c}{n^2} \quad (19)$$

DBO-EIF算法的基本流程为:①初始化DBO参数,将个体位置作为EIF参数进行训练以计算适应度值;②根据蛱螂的不同行为模式更新个体位置,并重新评估适应度,每次迭代中保留适应度更优的位置;③该过程循环执行直至满足终止条件,最终输出全局最优适应度对应的位置,即EIF最佳参数。图2为DBO-EIF算法流程图。

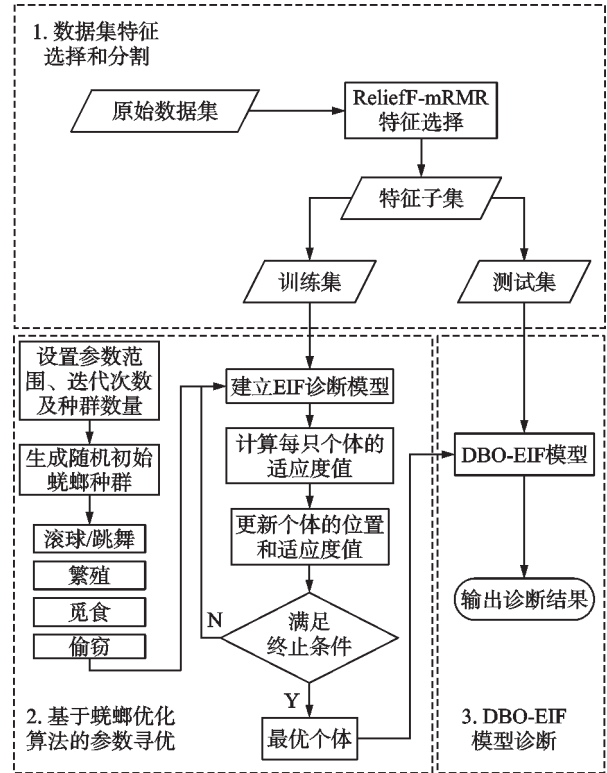


图2 DBO-EIF算法流程图

Fig.2 DBO-EIF algorithm flowchart

2 实验与分析

2.1 实验数据选取和特征选择

2.1.1 实验环境

实验计算机配备Intel Core i5-12500处理器,内存为16 GB,64位Windows11操作系统,实验代码基于Python3.9编写。

2.1.2 实验数据获取

由于大型望远镜驱动系统成本高昂且无法直接进行故障植入,故本研究采用半实物仿真平台进行数据采集与故障模拟。该平台模拟了望远镜的实际运行机制,采用位置环、速度环和电流三环闭环控制,由运动控制器、伺服驱动器、伺服电机及高精度位置传感器等构成。核心硬件包括:UMAC多轴运动控制器(Delta Tau Data Systems)、AKD-B01206伺服驱动器(Kollmorgen)以及AKM51H同步伺服电机(Kollmorgen)。伺服电机同轴连接一金属圆盘,用于支撑并定位29位绝对式光栅编码器。故障诊断及自愈半实物仿真平台如图3所示。

为了模拟不同层次的故障情况,使用压力传感器和螺丝阀门对驱动系统施加不同程度的摩擦力。将压力传感器通过可旋转螺杆固定在电机转台支撑架上,通过手动或自动的方式调节螺丝阀门的松紧来控制压力传感器与转台圆盘之间的压力,进而控制摩擦力的大小,达到调节伺服电机负载的目的。

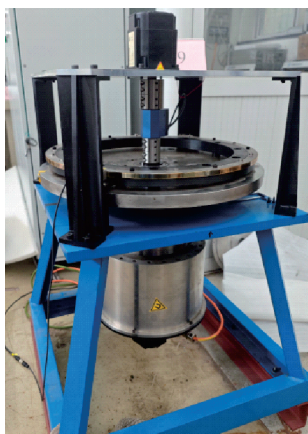


图3 故障诊断及自愈半实物仿真平台

Fig.3 Fault diagnosis and self-healing semi physical simulation platform

通过压力传感器读取负载压力的数值,实现波动负载、递增负载和递增后稳定这3种故障模式,通过USB通讯接口采集压力传感器数据并导入上位机。

望远镜的运行状态根据其在运行过程中产生的各项参数分析得到。在半实物仿真实验中,以望远镜常用的指向速度 $3\ 600\ (^{\circ})/\text{s}$ 运行驱动系统,对驱动系统的主要运行指标进行数据采集,包括实际位置、目标位置、实际速度、目标速度、实际加速度、目标加速度、跟随误差、三相电流电压和励磁电压,共14个特征。非预期故障诊断实验数据如表1所示。

表1 非预期故障诊断实验数据

Tab.1 Unanticipated fault diagnosis experimental data

编号	特征	编号	特征
S ₁	实际位置	S ₈	A相测量
S ₂	目标位置	S ₉	B相测量
S ₃	实际速度	S ₁₀	C相测量
S ₄	目标速度	S ₁₁	A相电压
S ₅	实际加速度	S ₁₂	B相电压
S ₆	目标加速度	S ₁₃	C相电压
S ₇	跟随误差	S ₁₄	D相电压

实验所用数据集选取空载运行数据2 700组,分别选取波动负载、递增负载和递增后稳定的3种故障模式100组,共3 000组实验数据,组成了 $3\ 000 \times 14$ 维的矩阵。将其中的空载运行数据作为正常数据,负载数据作为异常数据,即异常数据占比为5%,训练集与测试集的比例为7:3。在故障检测方法方面,使用了扩展孤立森林算法对训练数据进行学习。

2.1.3 特征选择

使用ReliefF-MRMR特征选择算法对原始数据集进行特征权重计算以及特征选择。首先,将原始数据集中的14维特征数据作为ReliefF算法的输

入,计算各个特征的权重,ReliefF算法的特征权重排序如图4所示;其次,设权重阈值为0.2,保留权重大于0.2的11维特征作为mRMR算法的输入,对剩余特征子集进一步优化,得到与目标类别有最大相关性且相互之间具有最小冗余性的特征重要度排序。mRMR算法的特征重要度排序如图5所示。实验中剔除重要度最小的2维,保留前9维特征进行后续模型训练。

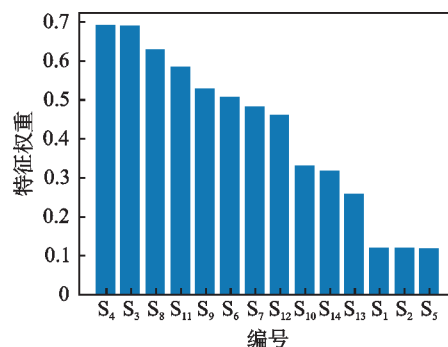


图4 ReliefF算法的特征权重排序

Fig.4 ReliefF feature weight sorting

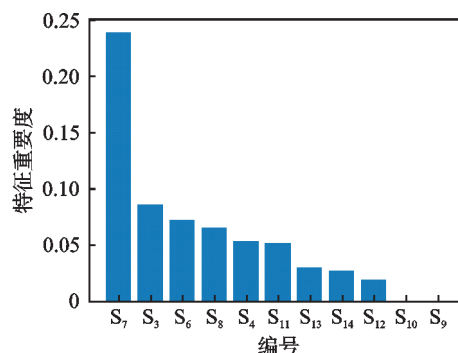


图5 mRMR算法的特征重要度排序

Fig.5 MRMR feature importance ranking

2.2 DBO-EIF 算法性能评估

为了验证所提出的DBO-EIF算法和特征选择算法的有效性,将经过ReliefF-mRMR特征选择算法优选前后的数据分别输入到DBO-EIF、EIF、IF、LOF、OCSVM和深度SVDD等异常检测模型中进行训练。训练完成后在测试集上输出故障诊断结果,DBO-EIF与其他异常检测算法性能对比如表2所示。其中,Rm表示采用ReliefF-mRMR特征选择算法选择后的特征子集。采用准确率、F1值、曲线下面积(area under the curve,简称AUC)值和Kappa系数综合评价模型效果,能够直观地表示模型优劣。

初始化DBO参数时,设置最大迭代次数为50,种群数量为30,孤立树数量范围为[1,500],子采样范围为[10,512],树高为[3,10],扩展等级为[1,8]。训练EIF并进行测试,图6为采用DBO算法优化EIF迭代过程。可见,DBO算法在迭代至17次时,

表 2 DBO-EIF 算法与其他异常检测算法性能对比
Tab.2 Performance comparison between DBO-EIF
and other anomaly detection algorithms

算法	准确率/ %	F1/ %	AUC	Kappa 系数	t/s
Rm-DBO-EIF	98.7	99.3	0.913	0.846	0.211
Rm-EIF	97.6	98.3	0.896	0.757	0.204
EIF	96.4	98.3	0.882	0.682	0.224
iForest	96.7	97.9	0.841	0.649	0.095
LOF	86.9	91.8	0.944	0.593	0.049
OCSVM	81.1	87.8	0.803	0.447	0.041
深度 SVDD	91.7	95.1	0.751	0.665	10.082

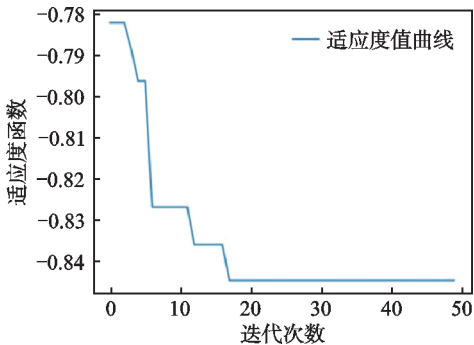


图 6 DBO 算法优化 EIF 迭代过程

Fig.6 The iterative process of optimizing EIF with DBO algorithm

适应度函数达到最小。经过 DBO 算法优化后的 EIF 最优参数值如下:孤立树个数为 285、子采样大小为 353、孤立树最大深度为 5、扩展等级为 6。

由表 2 可见,经过 DBO 算法优化后的 EIF 算法总体性能高于其他异常检测算法。具体来说,经过 DBO 优化后的 EIF 算法准确率从 97.6% 提升到 98.7%, F1 值从 98.3% 提升到 99.3%, AUC 值从 0.896 提升到 0.913, Kappa 系数从 0.757 提升到 0.846。Kappa 系数相比优化前提升约 11.7%,这表明算法对于数据量较小的故障数据的识别敏感度提升明显。

图 7 为不同异常检测算法的工作特征(receiver operating characteristic, 简称 ROC) 曲线。可见, DBO-EIF、EIF、IF 和 LOF 算法均能较好地识别数据集中的故障数据,具有较高的 AUC 值。具体来说, DBO-EIF 的 AUC 值优于除 LOF 之外的其他异常检测算法,虽然 LOF 算法的 ROC 曲线具有更高的 AUC 值,但是其准确率、F1 值和 Kappa 系数都较低,表明其对于数据量较小的故障数据敏感度较低,不适应本研究场景。

OCSVM、SVDD 和扩展孤立森林算法是 3 种基于切分超平面的异常检测方法。OCSVM 的目标是找到一个超平面,使其距离超平面最近的正常样本之间的距离最大化。SVDD 的目标是找到一个能将

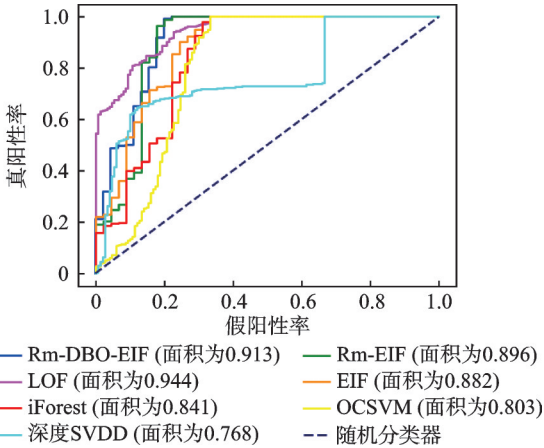


图 7 不同异常检测算法的 ROC 曲线

Fig.7 ROC curves of different algorithms

给定的所有正常数据包含在其中的最小半径的超球体。扩展孤立森林算法则需要找到一个使得划分后的 2 个子集中的样本数最小化的超平面。由表 2 可见, OCSVM 的预测准确率较低, Kappa 系数仅为 0.447, 故障样本敏感度较差。深度 SVDD 算法融入了神经网络的方法, 本实验中设训练轮数为 100, 模型训练时间较长。深度 SVDD 总体性能优于 OCSVM 和 LOF, 由图 7 可知, 其 ROC 曲线更接近图的右下方, AUC 值较低, 说明该算法对正常样本和故障样本区分度较差。

此外, 经过 ReliefF-mRMR 特征选择算法筛选后的特征子集在准确率、F1 值、AUC 值和 Kappa 系数这 4 个模型性能指标都优于原始数据集, 模型运行时间缩短了约 8.9%。实验表明, 此种特征选择算法能够有效地对原始数据集特征进行组合优化, 对模型的故障诊断性能和效率都有明显改善。

3 结束语

提出了一种基于 EIF 的非预期故障诊断方法。该方法将特征选择算法 ReliefF 与 mRMR 相结合, 构造出一个与目标类别有最大相关性且相互之间具有最小冗余性的特征子集, 有效改善了 EIF 对高维数据处理能力较弱的问题, 提高了诊断精度, 降低了训练时间。采用 DBO 算法对 EIF 算法进行参数优化, 得到最适合的非预期故障诊断模型, 在准确率、Kappa 系数等性能指标上均优于 LOF、OCSVM 和 SVDD 算法。实验表明, 该方法在大型望远镜驱动系统的非预期故障检测领域具有一定的可行性, 对提升望远镜驱动系统的可靠性和驱动系统故障的快速诊断有实质性意义。此外, 相较于原始孤立森林算法, 所提出算法分别对特征数据和模型参数进行优选, 可以构建出更加适配具体应用领域的模型, 具有广泛的应用前景。

参 考 文 献

- [1] 何章鸣. 非预期故障的数据驱动诊断方法研究[D]. 长沙: 国防科学技术大学, 2015.
- [2] 邓永停. 4m级望远镜主轴交流伺服控制系统研究[D]. 长春: 中国科学院研究生院(长春光学精密机械与物理研究所), 2015.
- [3] 邓壮壮, 杨世海, 朱节中, 等. 基于SVDD优化的南极望远镜驱动非预期故障研究[J]. 自动化与仪器仪表, 2022(1): 18-22.
DENG Zhuangzhuang, YANG Shihai, ZHU Jiezhong, et al. Unanticipated fault of driving antarctic telescope based on SVDD optimization[J]. Automation & Instrumentation, 2022(1): 18-22. (in Chinese)
- [4] PIMENTEL M A F, CLIFTON D A, CLIFTON L, et al. A review of novelty detection[J]. Signal Processing, 2014, 99: 215-249.
- [5] JIANG J C, LI T Y, CHANG C, et al. Fault diagnosis method for lithium-ion batteries in electric vehicles based on isolated forest algorithm[J]. Journal of Energy Storage, 2022, 50: 104177.
- [6] 余翔, 陈国洪, 李霆, 等. 基于孤立森林算法的用电数据异常检测研究[J]. 信息技术, 2018, 42(12): 88-92.
YU Xiang, CHEN Guohong, LI Ting, et al. Research on abnormal detection of electricity data based on isolated forest algorithm[J]. Information Technology, 2018, 42(12): 88-92. (in Chinese)
- [7] ALAZZAM H, SHARIEH A, SABRI K E. A lightweight intelligent network intrusion detection system using OCSVM and pigeon inspired optimizer[J]. Applied Intelligence, 2022, 52(4): 3527-3544.
- [8] KUMARAVEL V P, BUIATTI M, PARISE E, et al. Adaptable and robust EEG bad channel detection using local outlier factor (LOF) [J]. Sensors, 2022, 22(19): 7314.
- [9] WU X Y, LIU S Y, BAI Y G. The manifold regularized SVDD for noisy label detection[J]. Information Sciences, 2023, 619: 235-248.
- [10] 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. (in Chinese)
- [11] 李运, 杨世海, 吴佳佳, 等. 基于SVM观测器的新异类故障检测方法及应用[J]. 振动、测试与诊断, 2021, 41(2): 292-298.
LI Yun, YANG Shihai, WU Jiajia, et al. New outlier fault detection method and application based on SVM observer[J]. Journal of Vibration, Measurement & Diagnosis, 2021, 41(2): 292-298. (in Chinese)
- [12] 蔡炀烁, 高鹏骐, 沈鸣, 等. APOSOS项目15 cm光电望远镜故障监测方法研究[J]. 天文研究与技术, 2018, 15(2): 188-194.
CAI Yangshuo, GAO Pengqi, SHEN Ming, et al. Study on monitor and diagnostic method of APOSOS 15 cm opto-electrical telescopes[J]. Astronomical Research & Technology, 2018, 15(2): 188-194. (in Chinese)
- [13] LIU F T, TING K M, ZHOU Z H. Isolation forest [C]//2008 Eighth IEEE International Conference on Data Mining. Pisa, Italy: IEEE, 2008: 413-422.
- [14] MATTIA C, MATTEO T, ANTONIO S G. Interpretable anomaly detection with DIFFI: depth-based feature importance of isolation forest[J]. Engineering Applications of Artificial Intelligence, 2023, 119: 105730.
- [15] 杨先圣, 姜磊, 彭雄, 等. 基于大数据的异常检测方法研究[J]. 计算机工程与科学, 2018, 40(7): 1180-1186.
YANG Xiansheng, JIANG Lei, PENG Xiong, et al. A new outlier detection method based on large data[J]. Computer Engineering and Science, 2018, 40(7): 1180-1186. (in Chinese)
- [16] HARIRI S, KIND M C, BRUNNER R J. Extended isolation forest [J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 33(4): 1479-1489.
- [17] 孙林, 陈雨生, 徐久成. 基于改进ReliefF的多标记特征选择算法[J]. 山东大学学报(理学版), 2022, 57(4): 1-11.
SUN Lin, CHEN Yusheng, XU Jiucheng. Multilabel feature selection algorithm based on improved ReliefF [J]. Journal of Shandong University (Natural Science), 2022, 57(4): 1-11. (in Chinese)
- [18] ZHOU M R, BIAN K, HU F, et al. A new strategy for identification of coal miners with abnormal physical signs based on EN-mRMR[J]. Frontiers in Bioengineering and Biotechnology, 2022, 10: 935481.
- [19] XUE J K, SHEN B. Dung beetle optimizer: a new meta-heuristic algorithm for global optimization[J]. The Journal of Supercomputing, 2023, 79(7): 7305-7336.



第一作者简介:刘睿强,男,1999年5月生,硕士生。主要研究方向为天文仪器智能化故障诊断、机器学习算法。
E-mail: rqliu2021@niaot.ac.cn

通信作者简介:徐灵哲,男,1975年1月生,博士、副研究员、硕士生导师。主要研究方向为望远镜智能控制、望远镜观测控制系统和远程控制系统开发。
E-mail: lzhu@niaot.ac.cn