文章编号：1003－1251(2024)05－0041－08

基于可解释性 USL 的抽油井电功率异常检测

袁春华，冯博文，李翔宇

（沈阳理工大学 自动化与电气工程学院，沈阳 110159）

摘　要：抽油井长期运行可能使有杆泵抽油系统出现异常工况，这种异常的产生与设备的持续退化有关，但其具体原因仍难以通过数据驱动的深度学习方法来分析和解释。为了能够及时准确地检测到异常工况并推断故障级别和故障类型，提出一种基于可解释性无监督 Shapelet 学习（unsupervised Shapelet learning，USL）算法的抽油井电功率异常检测方法。首先利用提取到的先验知识改进 USL；然后将改进的 USL 与单分类器结合，构成基于可解释性 USL 的异常检测算法（AD-IUSL），AD-IUSL 可同时学习 Shapelet 特征和异常阈值。利用具有不同异常工况的抽油井电功率数据进行验证，结果表明，本文提出算法的 F1 分数达到了 0.969，根据学习到的 Shapelet 特征合理解释了两种异常发生的潜在原因。

关　键　词：抽油井；异常检测；可解释性；Shapelet 学习算法

中图分类号：TP277.3　　**文献标志码：**A　**DOI：**10.3969/j.issn.1003－1251.2024.05.007

Anomaly Detection of Pumping Well Motor Data Based on Interpretable USL

YUAN Chunhua，FENG Bowen，LI Xiangyu

（Shenyang Ligong University，Shenyang 110159，China）

Abstract： Long-term operation of oil wells can lead to abnormal conditions in the sucker rod pump system. These abnormalities are associated with the continuous degradation of equipment. However，it is still difficult to analyse and interpret the specific causes using data-driven deep learning methods. To detect such abnormal conditions promptly and accurately and to deduce the level and type of faults，an anomaly detection method for the motor power of oil wells is proposed based on an interpretable unsupervised Shapelet learning （USL） algorithm. Initially，the USL is improved using extracted prior knowledge. Subsequently，the USL is combined with a one-class classifier to form an anomaly detection algorithm based on interpretable USL（AD-IUSL），which can simultaneously learn Shapelets and anomaly thresholds. This study validates the proposed method using electrical power data from oil wells with two different abnormal conditions. Experimental results demonstrate that the proposed algorithm achieves an F1-score of 0.969 and can explain the potential reasons for two types of anomalies based on the learned Shapelet feature.

Key words： pumping well；anomaly detection；interpretability；Shapelet learning algorithm

收稿日期：2023－10－08

基金项目：国家自然科学基金项目（62173073）；辽宁省教育厅高等学校基本科研项目（LJKMZ20220618，JYTMS20230215）；辽宁省本科教改优质教学资源建设与共享项目（SBKJGYZ－2021－06）

作者简介：袁春华（1987—），女，副教授，博士，研究方向为复杂工业过程建模、诊断与控制；李翔宇（1982—），通信作者，男，副教授，博士，研究方向为复杂工业过程建模、诊断与控制。

有杆泵抽油系统是陆上油田的重要组成部分,用于从地下油藏中汲取原油,系统出现故障会停井维修,增加运维成本。因此,及时、准确地诊断系统的故障至关重要。基于示功图的计算机辅助故障诊断[1]是一种降低上述风险的传统方法,但该方法存在故障检测延迟、安装传感器成本高等问题[2]。

目前,故障诊断已经转向使用电机功率数据,该方法的优点是可连续监测、数据采集高频、易于维护[3-4],此外电功率具有时间序列的性质,蕴含了丰富的信息,但现阶段电功率数据标签稀缺[5],且对其特征提取依赖于完备的机理模型[6-7]。为解决上述问题,往往采用基于深度学习的时间序列异常检测模型。时间序列异常可分为时间点之间、时间间隔之间、时间序列之间异常三种类型[8],目前能够同时解决这三类问题的方法是基于长短时记忆网络(LSTM)框架下的数据驱动模型[9-10]。数据驱动方法避免了过于依赖先验知识的问题,但深度学习不可避免地存在可解释性差的问题[11],即很难理解检测到的异常发生的原因。

Shapelet 学习算法具有准确性高、可解释性强、时间复杂度较低等优点[12],是目前用于时间序列分类和异常检测的重要方法。传统的 Shapelet 学习依赖于有标签数据集的监督学习[13],为了应用于时间序列异常检测,其被改进为无监督 Shapelet 学习(unsupervised Shapelet learning, USL)算法[14],但该算法在工业领域应用中由于受到噪声的影响,会学习到与特定领域无关的特征,从而导致误分类。

为解决上述问题,本文提出一种基于可解释性无监督 Shapelet 学习的异常检测(anomaly detection with interpretable unsupervised Shapelet learning, AD-IUSL)算法,用于抽油井电功率异常检测。首先利用先验知识改进 USL;然后将改进的算法与单分类器结合,形成端到端的异常检测算法 AD-IUSL;最后利用 AD-IUSL 对抽油井电功率长期数据进行检测,解释异常发生的可能原因。

1 无监督 Shapelet 学习算法

1.1 传统无监督 Shapelet 学习算法

传统 USL 用于时间序列分类问题时包括 Shapelet 生成、Shapelet 转换和 Shapelet 分类三个阶段[15],且上一阶段的结果作为下一阶段的输入。为将 Shapelet 学习算法应用到异常检测中,利用无监督数值优化主成分分析(principal components analysis, PCA)方法生成 Shapelet。

PCA 的数学原理是最大化投影数据的方差,找到最大特征值对应的特征向量。Shapelet 特征的关键属性为“形状”和位置,其中“形状”由每个数据点的值决定。因此,在一个时间序列数据集中,利用特征向量指向 Shapelet 的位置并生成 Shapelet 的表达式为

$$\begin{aligned} \boldsymbol{v}^{*} &= \underset{\boldsymbol{v}}{\operatorname{argmin}}\ \boldsymbol{v}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{v} \\ &\text{s.t.} \quad \boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}=1 \end{aligned} \tag{1}$$

式中:$\boldsymbol{v}^{*}$ 为目标特征向量;$\boldsymbol{v}$ 为特征向量;$\boldsymbol{C}$ 为时间序列数据集中所有样本的协方差矩阵。

1.2 稀疏模型

为求解公式(1),常用的方法是引入正则化项,使解具有稀疏性,一般采用最小绝对收缩和选择算子(LASSO)算法。将式(1)改写为

$$\begin{aligned} \boldsymbol{v}^{*} &= \underset{\boldsymbol{v}}{\operatorname{argmin}}\ \boldsymbol{v}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{v}+\alpha_1 \| \boldsymbol{v} \|_1 \\ &\text{s.t.} \quad \boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}=1 \end{aligned} \tag{2}$$

式中:α_1 为正则化项系数;$\| \boldsymbol{v} \|_1$ 表示特征向量的 L1 范数。由于式(2)具有较弱的块状结构,LASSO 算法依然不利于将优化问题分解为子优化问题,并利用增广拉格朗日方程等方法求解。

2 先验知识提取及其数学表示

2.1 有杆泵抽油系统机理模型

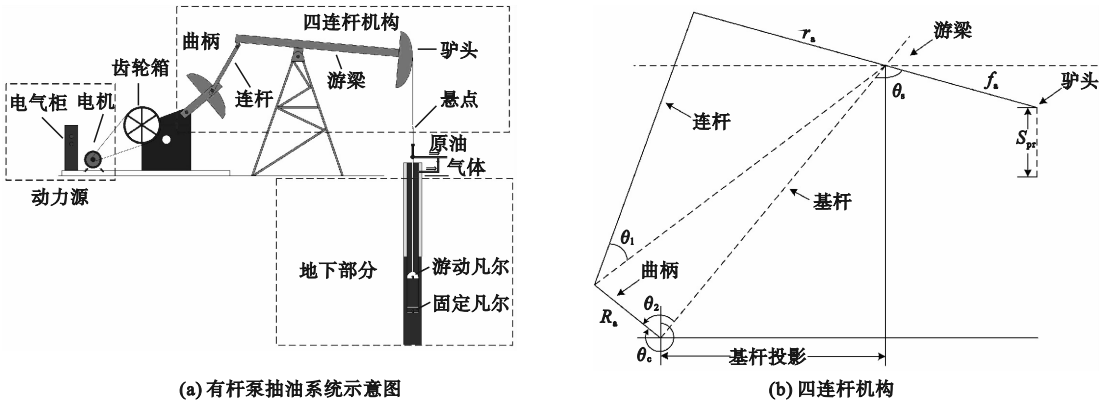
有杆泵抽油系统机理模型如图 1 所示。有杆泵抽油系统由动力源、四连杆机构(如图 1(b)所示)和地下部分组成。主要工作过程为:动力源由电网侧供电,使感应电机工作,齿轮箱将高转速低扭矩转换为低转速高扭矩,驱动曲柄做圆周运动,从而使四连杆机构中的悬点做往复运动。

动力源和四连杆机构中的部分变量如表 1 所示,这些变量分为三类,分别表示对系统的不同影响。图 1(b)中所示的其他变量在下文中给出说明。

1)输入控制变量:影响系统的外部因素。

2)状态变量:描述系统的基本变量,隐式地表示系统如何运行。

3)系统决策变量:由状态变量或由机械结构参数决定的变量,显式地决定系统运行动态。



(a) 有杆泵抽油系统示意图　　(b) 四连杆机构

图 1　有杆泵抽油系统机理模型

Fig. 1　Mechanism model of sucker rod pump system

表 1　有杆泵抽油系统的部分变量

Table 1　Variables in the sucker rod pump system

变量类别	变量名称
输入控制变量	电压 $\boldsymbol{U}$ 负载扭矩 T_{L}
状态变量	电流 $\boldsymbol{I}$ 转子角位移 Θ 转子角速度 Ω_{r} 游梁和基杆之间的夹角 θ_{s} 曲柄角位移 θ_{c}
系统决策变量	电磁扭矩 T_{e} 悬点位移 S_{pr} 曲柄驱动扭矩 T_{C} 曲柄阻力扭矩 T_{D} 扭矩因数 $\overline{TF}$

2.2　先验知识提取

基于上述系统决策变量中最能表征电机功率特性和系统运动规律的最小系统决策变量组提取先验知识,融入到 Shapelet 学习算法中。本文通过感应电机和四连杆机构的运动学及动力学模型建立变量间的联系,实现先验知识的提取。

1)感应电机

电磁扭矩为系统感应电机输出的驱动力矩。由于本节旨在先验知识的提取而非仿真电机功率曲线,因此不考虑铁损耗、铜损耗,此时电磁扭矩的表达式为

$$T_{\mathrm{e}}=\frac{1}{2}p_0\boldsymbol{I}^{\mathrm{T}}\frac{\partial\boldsymbol{L}}{\partial\Theta}\boldsymbol{I} \tag{3}$$

式中:p_0 为感应电机的极对数;$\boldsymbol{I}$ 是由定子电流和转子电流构成的矩阵;$\boldsymbol{L}$ 为电机电感矩阵。

2)四连杆机构

将曲柄轴的圆周运动转化为悬点的往复运动,驱动力矩为经齿轮箱转换后的电磁扭矩,不考虑平衡重滞后角度带来的影响,曲柄轴运动方程[15]为

$$\begin{cases}J_{\mathrm{e}}\dfrac{\mathrm{d}^2\theta_{\mathrm{c}}}{\mathrm{d}t^2}+\dfrac{1}{2}\dfrac{\mathrm{d}\theta_{\mathrm{c}}}{\mathrm{d}t}\dfrac{\mathrm{d}J_{\mathrm{e}}}{\mathrm{d}\theta_{\mathrm{c}}}=T_{\mathrm{C}}-T_{\mathrm{D}}\\ T_{\mathrm{C}}=\eta k\cdot T_{\mathrm{e}}\\ T_{\mathrm{D}}=\overline{TF}\cdot T_{\mathrm{EL}}+\sin\theta_{\mathrm{c}}\cdot T_{\max}\end{cases} \tag{4}$$

式中:t 为时间;J_{e} 为系统等效惯量;η 和 k 分别为感应电机和齿轮箱的传动效率;T_{EL} 为曲柄轴到驴头的等效扭矩;$T_{\max}$ 为最大平衡扭矩。扭矩因数仅由系统的机械参数决定,随着机械角度的变化而变化,可表示为

$$\overline{TF}=R_{\mathrm{a}}\frac{f_{\mathrm{a}}}{r_{\mathrm{a}}}\cdot\frac{\sin(\theta_2-\theta_1)}{\sin(\theta_1-\theta_{\mathrm{s}})} \tag{5}$$

式中:R_{a} 为曲柄半径;f_{a} 和 r_{a} 分别为游梁前臂和后臂长度;θ_1 和 θ_2 如图 1(b)所示。

由式(4)可知,决定系统驱动扭矩及系统阻力扭矩的关键变量为电磁转矩、扭矩因数和曲柄角位移,结合电机功率 P 的如下表达式

$$P=T_{\mathrm{e}}\cdot\Omega_{\mathrm{r}} \tag{6}$$

$$P=\overline{TF}\cdot L_{\mathrm{EL}}\cdot\Omega_{\mathrm{c}} \tag{7}$$

式中:L_{EL} 为 T_{EL} 对应的等效负载;Ω_{c} 为曲柄角速度。可知,T_{e} 和 $\overline{TF}$ 直接影响电机功率特性。

此外,为捕捉游动凡尔、固定凡尔打开和闭合四个关键工作节点的信息,将悬点位移考虑为表示系统运动规律的变量,表达式为

$$S_{\mathrm{pr}}=f_{\mathrm{a}}\cdot\theta_{\mathrm{s}} \tag{8}$$

$$\theta_{\mathrm{s}}=\phi_\theta(\theta_{\mathrm{c}}) \tag{9}$$

式中 ϕ_θ 表示 θ_{s} 与 θ_{c} 之间关系的非线性函数。

综上,选择 T_{e}、$\overline{TF}$ 和 S_{pr} 为最小系统决策变量组,三个变量的仿真曲线如图 2 所示。

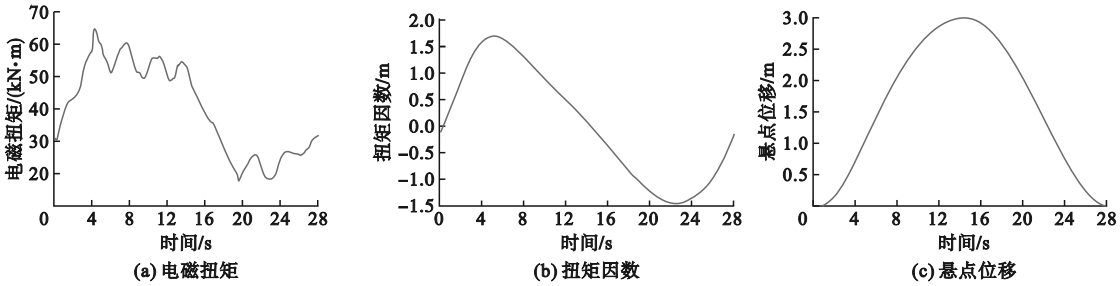


图 2 最小决策变量组仿真曲线

Fig. 2 Minimum decision variable group simulation curves

2.3 先验知识数学表示

式(2)中较弱的块状结构问题通过再次引入一个正则化项来解决。最小系统决策变量组以先验知识矩阵的方式加入正则化项中,将式(2)改写为

$$\boldsymbol{v}^* = \underset{\boldsymbol{v}}{\operatorname{argmin}}\ \boldsymbol{v}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{v} + \alpha_1 \| \boldsymbol{v} \|_1 + \alpha_2 \| \boldsymbol{D}\boldsymbol{X}\boldsymbol{v} \|_1$$
$$\text{s.t.} \quad \boldsymbol{v}^{\mathrm{T}}\boldsymbol{v} = 1 \tag{10}$$

式中:α_2 为正则化项系数;$\boldsymbol{X} \in \mathbf{R}^{N\times M}$,为时间序列数据集,其中 N 为时间序列数据集中时间序列的数量,M 为每个时间序列的数据点个数;$\boldsymbol{D} \in \mathbf{R}^{V\times N}$,为先验知识矩阵,其中 V 为最小系统决策变量的数量。根据式(9),利用先验矩阵对应于时间序列的方式捕捉时间序列中关键的工作节点,并学习到能包含这些节点的 Shapelet 特征向量 $\boldsymbol{v}$,最终由目标特征向量 $\boldsymbol{v}^*$ 指示的非零块状结构得到 Shapelet,表达式为

$$\boldsymbol{\sigma} = \mathrm{Ind}(\boldsymbol{v}^*) \tag{11}$$

式中:$\boldsymbol{\sigma}$ 为学习到的 Shapelet;Ind(·)表示指示函数,当满足条件时赋值为 1,否则赋值为 0。

3 基于可解释性 USL 异常检测算法

将 USL 融入异常检测中,首先要定义本文提出的 Shapelet 转换机制[16],其次提出 AD-IUSL 的全局目标函数,同时学习到具有可解释性的 Shapelet 特征和异常阈值,最后利用增广拉格朗日方程求解此目标函数。

3.1 Shapelet 转换机制

Shapelet 转换机制本质上是一种基于距离定义的相似性度量,对于给定长度为 l 的某时间序列的 Shapelet,转换机制表达式为

$$\phi_\sigma(\sigma_k, X_j) = \sum_{j=k}^{k+l-1} \frac{1}{l} \| X_j - \sigma_k \|_2^2 \tag{12}$$

式中:$\phi_\sigma(\cdot)$表示 Shapelet 的转换函数;X_j 表示某时间序列的第 j 个数据点;σ_k 表示 Shapelet 的起始数据点。将式(12)进一步推广到有 Q 个不同长度 l_q($q = 1,2,\cdots,Q$)的 Shapelet 集,此时转换函数表达的矩阵 $\boldsymbol{\Phi}_\sigma$($\boldsymbol{\Phi}_\sigma \in \mathbf{R}^{N\times Q}$)可表示为

$$\boldsymbol{\Phi}_\sigma(\boldsymbol{\sigma}, \boldsymbol{X}) = \underset{\substack{q=1,\cdots,Q\\ i=1,\cdots,N}}{\phi_\sigma}(\sigma_{k_q,l_q}, X_{i,j}) \tag{13}$$

式中:σ_{k_q,l_q}表示长度为 l_q 且数据起始点为 k_q 的第 q 个 Shapelet;$X_{i,j}$表示第 i 个时间序列 $\boldsymbol{X}_i$ 的第 j 个数据点。

3.2 目标函数

在定义转换机制的基础上,AD-IUSL 的目标函数需要将转换机制下的转换矩阵整合到分类器中。转换矩阵 $\boldsymbol{\Phi}_\sigma$ 中的每个元素在单分类器中是降维后的点到原点的距离,在异常检测中亦可称为异常分数。异常分数 r_{iq}可简化表示为

$$r_{iq} = \phi_\sigma(\sigma_{k_q,l_q}, X_{i,j}) \tag{14}$$

由于异常检测需要单分类器有一个可被优化的阈值,因此 AD-IUSL 引入具有软阈值的支持向量数据描述(SVDD)分类器,其目标函数 F 表达式为

$$F(\sigma_q^*, R^*) = \underset{\sigma_q, R}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{q=1}^{Q} \boldsymbol{\Phi}_\sigma(\boldsymbol{\sigma}, \boldsymbol{X}) + R^2 + B \cdot \sum_{i=1}^{N} \max\{0, r_{iq} - R^2\} \tag{15}$$

式中:σ_q^* 为目标 Shapelet;R^* 为目标异常阈值;R 为异常阈值;B 为正则化参数,用于平衡假正样本与假负样本之间的比率。

3.3 目标函数优化

由于 $\boldsymbol{v}$ 和 R 的存在,式(15)是一个非凸优化问题。本文采用组合优化的方式交替优化 USL 部分和单分类部分。

3.3.1 USL 优化

当更新 Shapelet 时固定阈值,此时式(15)表示为

$$F(\sigma_q^*) = \underset{\sigma_q}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{q=1}^{Q} \boldsymbol{\Phi}_\sigma(\boldsymbol{\sigma}, \boldsymbol{X}) \tag{16}$$

求解式(16)等同于优化式(10)。式(10)中

二次项的存在导致特征向量 $\boldsymbol{v}$ 并不满足并行更新的条件。因此,采用交替方向乘子法[17]将优化问题分解为子问题进行求解。

首先,在固定约束条件 $\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}=1$ 下,式(10)可表达为

$$\begin{aligned}&\min_{\boldsymbol{v},z_1,z_2}\boldsymbol{v}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{v}+\alpha_1\|z_1\|_1+\alpha_2\|z_2\|_1\\&\text{s.t.}\quad z_1=\boldsymbol{D}\boldsymbol{X}\boldsymbol{v},z_2=\boldsymbol{v}\end{aligned}\tag{17}$$

根据上式引入增广拉格朗日方程

$$\begin{aligned}L_\rho=\boldsymbol{v}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{v}+\boldsymbol{\lambda}_1^{(t)\mathrm{T}}(\boldsymbol{D}\boldsymbol{X}\boldsymbol{v}-z_1^{(t)})+\boldsymbol{\lambda}_2^{(t)\mathrm{T}}(\boldsymbol{v}-z_2^{(t)})+\\\frac{\rho_1}{2}\left\|\boldsymbol{D}\boldsymbol{X}\boldsymbol{v}-z_1^{(t)}\right\|_2^2+\frac{\rho_2}{2}\left\|\boldsymbol{v}-z_2^{(t)}\right\|_2^2\end{aligned}\tag{18}$$

式中:L_ρ 为增广拉格朗日方程函数;t 表示迭代次数;$\boldsymbol{\lambda}_1^{(t)}$、$\boldsymbol{\lambda}_2^{(t)}$ 为第 t 次迭代时的对偶变量;ρ_1、ρ_2 为正则化项参数。简化表示原始变量 $\boldsymbol{v}$ 的更新过程为

$$\boldsymbol{v}^{(t+1)}=\underset{\boldsymbol{v}}{\operatorname{argmin}}\,L_\rho(\boldsymbol{v},z_1^{(t)},z_2^{(t)},\boldsymbol{\lambda}_1^{(t)},\boldsymbol{\lambda}_2^{(t)})\tag{19}$$

利用矩阵求逆引理得到特征向量 $\boldsymbol{v}$ 为

$$\begin{aligned}&\boldsymbol{v}=(2\boldsymbol{C}+\rho_1(\boldsymbol{D}\boldsymbol{X})^{\mathrm{T}}\boldsymbol{D}\boldsymbol{X}+\rho_1\boldsymbol{E})^{-1}\\&(\rho_1(\boldsymbol{D}\boldsymbol{X})^{\mathrm{T}}z_1^{(t)}+\rho_2 z_2^{(t)}-(\boldsymbol{D}\boldsymbol{X})^{\mathrm{T}}\boldsymbol{\lambda}_1^{(t)}-\boldsymbol{\lambda}_2^{(t)})\end{aligned}\tag{20}$$

式中 $\boldsymbol{E}$ 表示单位矩阵。

其次,更新 z_1 和 z_2。当确定了 $\boldsymbol{v}$ 和 z_2 时,可以通过软阈值表示的方式更新 z_1。为表示简便,设 $\boldsymbol{\mu}_1^{(t)}=\boldsymbol{\lambda}_1^{(t)}/\rho_1$,$\boldsymbol{\mu}_2^{(t)}=\boldsymbol{\lambda}_2^{(t)}/\rho_2$。

$$z_1^{(t+1)}=S_{\alpha_1/\rho_1}(\boldsymbol{D}\boldsymbol{X}\boldsymbol{v}^{(t+1)}+\boldsymbol{\mu}_1^{(t)})\tag{21}$$

式中 $S_*(\cdot)$ 表示软阈值算子。$S_*(\cdot)$ 表达式为

$$S_*(\cdot)=\operatorname{sgn}(\cdot)\cdot\max(|\cdot|-*,0)\tag{22}$$

类似地,当确定了 $\boldsymbol{v}$ 和 z_1 时,按下式更新 z_2。

$$z_2^{(t+1)}=S_{\alpha_2/\rho_2}(\boldsymbol{v}^{(t+1)}+\boldsymbol{\mu}_2^{(t)})\tag{23}$$

3.3.2　阈值优化

根据 3.3.1 中方法固定 Shapelet 后,更新阈值,此时式(15)表示为

$$F(R^*)=\underset{R}{\operatorname{argmin}}\,R^2+B\cdot\sum_{i=1}^{N}\max\{0,r_{iq}-R^2\}\tag{24}$$

该式具有最大值函数项,对 R 不可微,故采用子梯度方法对其优化。R 子梯度计算式为

$$\frac{\partial F(R^*)}{\partial R}=\begin{cases}2R-2BN_{\mathrm{A}}R, & r_{iq}>R^2\\2R, & r_{iq}\leqslant R^2\end{cases}\tag{25}$$

式中 N_{A} 为异常样本个数,当 $r_{iq}>R^2$ 时出现异常,将梯度置 0,则 $N_{\mathrm{A}}=1/B$。

4　实验结果与分析

为验证本文提出算法的准确性、可解释性、有效性,利用辽宁某油田的两口故障程度不同(严重故障、次级故障)的抽油井中两个长时电功率数据集进行实验,数据集由安装在电气柜中的自主开发设备采集得到。

4.1　数据集描述

两个故障程度不同的数据集源自相同类型的有杆泵抽油系统(CYJ10-3-53-HB),分别为严重故障和次级故障。严重故障的数据集涵盖了为期一个月的电功率连续信号,其中在第 25 天后检测到了严重故障,数据集的单周期时间从 25 s 逐渐过渡到 53 s,相当于每个样本的数据点从 228 个逐渐变化到 478 个;次级故障的数据集涵盖了为期两个月的电功率连续信号,其中在第 45 天后发生了次级故障,数据集中的每个周期时间均为 28 s,相当于每个样本数据点为 250 个。两个数据集均以 8∶2 的比例划分为训练集和测试集。

4.2　定量比较

采用准确率与 F1 分数两种评估指标验证本文提出方法的准确性。准确率表示模型正确预测(包括正类和负类)的总数与总预测数的比例;F1 分数是一种结合了精确率和召回率的指标,常用于不平衡数据集的情况[9]。

将本文方法与六种不同的基线方法进行对比。六种基线方法中 OC-SVM 和 KNN 为基本的时间序列分类方法,ADSL[14] 和 LSTM-AE[9] 均为数据驱动的端到端异常检测方法,LTS + OC-SVM[16] 和 LSTM + OC-SVM[10] 均为两阶段(特征提取和工况识别)异常检测方法。各基线方法的实验超参数设置如表 2 所示。

表 2　各基线方法的超参数设置

Table 2　Hyperparameters configuration for various baseline methods

基线方法	超参数设置
OC-SVM	$\gamma=0.5$
KNN	$neighbor=5$
ADSL[14]	异常率 $\alpha=5\%$ Shapelet 数量 $Q=2$ Shapelet 长度 $l=0.08M$
LSTM-AE[9]	神经元数量为 128 隐含层状态向量维数为 64 反向传播时间序列长度为 32 $dropout=0.2$
LTS + OC-SVM[16]	Shapelet 数量 $Q=2$ Shapelet 长度 $l=0.08M$
LSTM + OC-SVM[10]	全连接层数量为 840 $\gamma=0.5$

实验结果如图 3 所示。可见,各方法的准确率均高于 80%,但除 LSTM-AE 和本文方法外,其余方法的 F1 分数均低于其准确率,这说明其余方法在不平衡数据集上表现较差。异常检测属于不平衡数据集下的典型情况,其异常样本远远少于正常样本。LSTM-AE 的 F1 分数达到 0.962,与本文方法的 F1 分数(0.969)几乎持平,但 LSTM-AE 作为一种深度学习方法,不仅对超参数设置要求严苛,且相较于本文方法可解释性差,使得技术人员无法判断异常出现的可能原因。

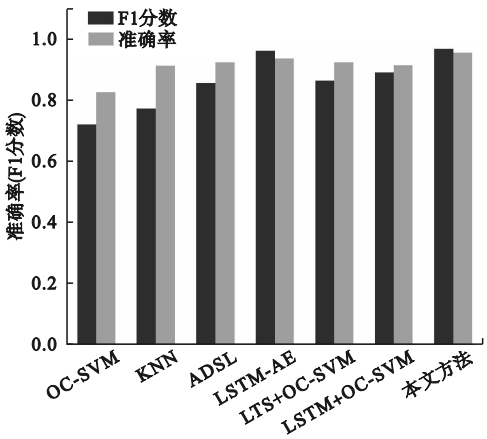


图 3 定量比较结果

Fig. 3 Results of quantitative comparison

4.3 Shapelet 可解释性分析

通过出现严重故障和出现次级故障的两个数据集分析 Shapelet 在融入知识后的可解释性和在不同工况下的可解释性。Shapelet 的如下两个属性共同决定了电功率出现异常的原因:Shapelet 与原时间序列间的距离;Shapelet 出现的位置。

4.3.1 案例 1

案例 1 中的 Shapelet 分析如图 4 所示,根据图 4 阐明改进 USL 的内在筛选机制。

1)候选子序列 1

该子序列匹配了第一个峰值出现时的位置,且与原时间序列的距离(d)为 0.315 9。由图 4(a)可知,其没有学习到这一期间内的电功率趋势。从机理模型角度看,候选子序列 1 错过了固定凡尔打开节点的信息。

2)候选子序列 2

该子序列的学习过程受到噪声的影响,与原时间序列的距离为 0.421 9。这一情况阐释了原始 USL 是一种仅受“形状”影响的数据驱动算法。

3)Shapelet(候选子序列 3)

该子序列被选择成为最终的 Shapelet,不仅因为这一子序列匹配了第二个峰值出现的位置,而且学习到了下冲程期间第二个峰值的趋势。此外,这一子序列与原时间序列具有距离最小值(0.172)。从机理模型角度,Shapelet 学习到了游动凡尔打开节点的信息。

4)候选子序列 4

与候选子序列 2 相似,这一子序列的学习过程依然受到了噪声的影响,与原时间序列的距离为 0.609 4,未能捕获到可解释性的信息。

依据图 4(a),通过值与位置两个属性解释异常出现的原因。由图 4(b)可见,相较于 Shapelet,异常工况导致电功率曲线中的第二个峰值出现了明显下降,但第二个峰值出现的位置与图 4(a)几乎相同。因此,推断这一数据集中出现的异常可能为抽油杆断脱。

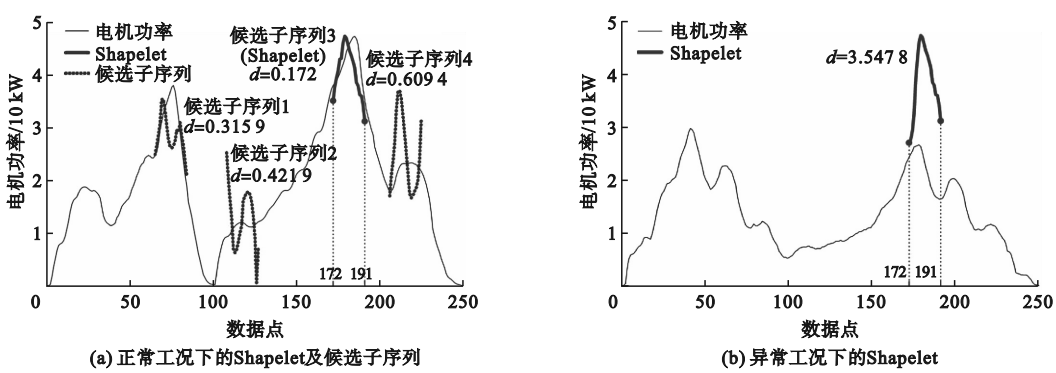


图 4 案例 1 中的 Shapelet 分析

Fig. 4 Shapelet analysis in case 1

4.3.2 案例 2

案例 2 中的 Shapelet 分析如图 5 所示。图 5(a)展示了 Shapelet 的两个属性,虽然粗点线的候选子序列已经学习到了曲线的趋势并具有距离最

小值,但并不能代表这一数据集内的电功率特征。图 5(b)中异常工况下电功率曲线的第一个峰值延迟出现,且其数值有所增加。因此,推断这一数据集中出现的异常可能为气体影响或固定凡尔漏失中的一种。

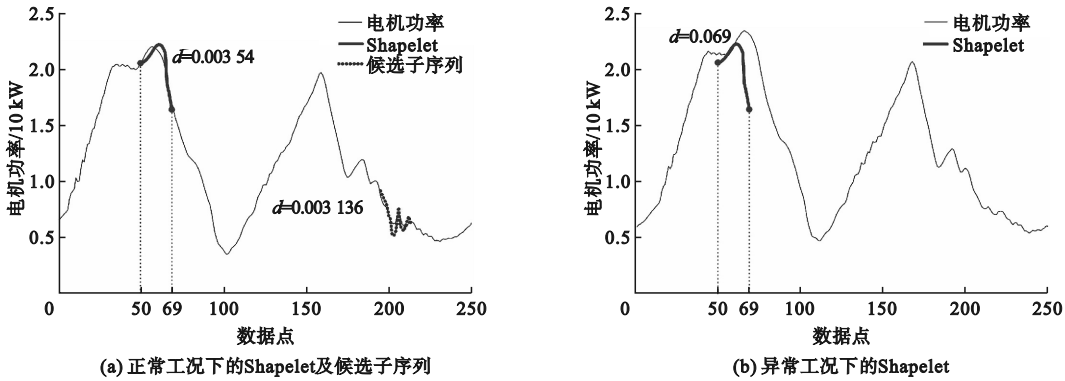


图 5　案例 2 中的 Shapelet 分析

Fig. 5　Shapelet analysis in case 2

4.4　工况变化趋势分析

4.4.1　严重故障

严重故障的工况分析如图 6 所示。由图 6(a)原始电功率数据可见,从第四个周期开始,第二个峰值出现了明显下降。由图 6(b)工况趋势可见,异常分数表现为短期内激增,由全局目标函数计算得到的阈值为 1.927 9。

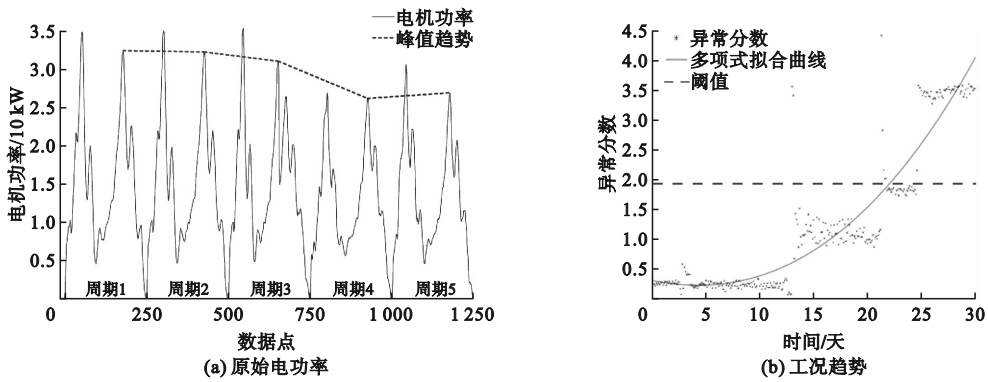


图 6　严重故障工况分析

Fig. 6　Severe fault working conditions analysis

从机理角度看,如果抽油杆断裂或脱离,游动凡尔可能不会及时打开,导致电功率第二个峰值明显下降或缺失。因此,推断此异常为严重故障,故障类型为抽油杆断脱。

4.4.2　次级故障

次级故障的工况分析如图 7 所示。由于次级故障的异常并非短期内造成,因此图 7(a)中每个周期第一个峰值没有明显变化。由图 7(b)可见,由全局目标函数计算得到的阈值为 0.066 79,异常分数缓慢增加,从第 45 天开始,异常分数频繁超过阈值且多项式曲线曲率明显增大。从机理角度将三个可能的故障分别说明如下。

1)气体影响

如果有大量气体进入泵腔中会导致泵效下降。这种情况往往导致固定凡尔延迟打开,峰值出现滞后。

2)固定凡尔漏失

如果固定凡尔不能正常打开,地层中的流体就无法进入泵筒,导致泵效低下,这种情况导致峰值出现滞后。

3)蜡沉积

对于油藏为稠油的油田,流体黏度随着蜡沉积而增大,油井产液量降低,生产压力梯度增加。这种情况也可能导致异常分数逐渐缓慢上升。

因此,推断该异常为次级故障,故障类型可能为气体影响、固定凡尔漏失或蜡沉积中的一种。

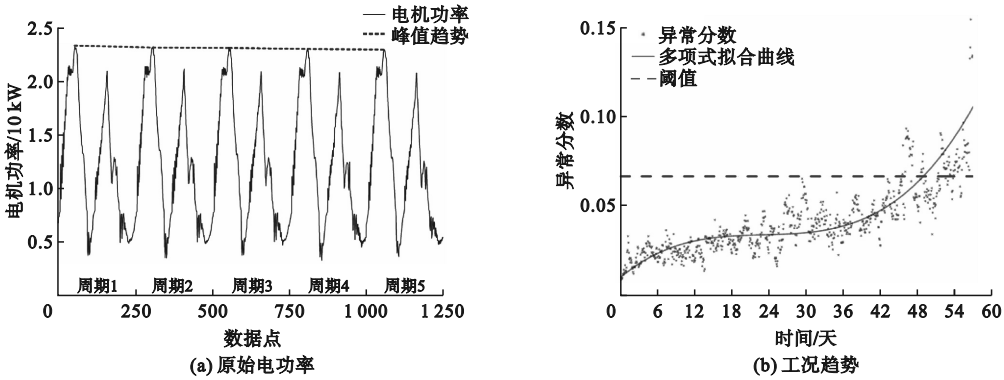


图 7 次级故障工况分析

Fig. 7 Secondary fault working conditions analysis

5 结论

本文首先通过将机理模型中的最小系统决策变量组形成先验知识矩阵并融入 USL,提取出最具可解释性的 Shapelet 特征;再将可解释性 USL 整合到异常检测中,形成了一个端到端的异常检测算法 AD-IUSL,该算法能够同时输出 Shapelet 特征和异常阈值;最后通过交替方向乘子法求解此算法的目标函数。利用辽宁某油田两口不同抽油井的长期电功率数据验证了本文算法的准确性、可解释性和有效性。本文算法的 F1 分数达到了 0.969,略优于目前最先进的深度学习异常检测算法;通过 Shapelet 的位置和"形状"两个固有属性,从故障类型和故障水平两个角度解释了系统出现异常的可能原因。

本文虽然采用 Shapelet 异常检测算法解释了电机功率数据出现异常的潜在原因,但并没有定量分析故障类型。今后将通过有监督学习在电功率数据基础上建立故障描述的先验知识库,并在小样本甚至零样本情况下进行具有可解释性的无监督异常检测。

参考文献(References):

[1] 李翔宇,袁春华,高宪文. 抽油井示功图诊断分析方法研究[J]. 沈阳理工大学学报,2021,40(1):88 – 94.
LI X Y, YUAN C H, GAO X W. Diagnosis of dynamometer card of pumping well[J]. Journal of Shenyang Ligong University, 2021, 40(1): 88 – 94. (in Chinese)

[2] ZHENG B Y, GAO X W, LI X Y. Diagnosis of sucker rod pump based on generating dynamometer cards[J]. Journal of Process Control, 2019, 77: 76 – 88.

[3] CHEN L, GAO X W, LI X Y. Using the motor power and XGBoost to diagnose working states of a sucker rod pump [J]. Journal of Petroleum Science and Engineering, 2021, 199: 108329.

[4] TECLE S I, ZIUZEV A, SPIRIDONOV V. Development of fault diagnosis model for sucker rod pump using motor power curve [C]//2021 XVIII International Scientific Technical Conference Alternating Current Electric Drives (ACED). Ekaterinburg, Russia: IEEE, 2021: 1 – 6.

[5] LUNDGREN A, JUNG D. Data-driven fault diagnosis analysis and open-set classification of time-series data[J]. Control Engineering Practice, 2022, 121: 105006.

[6] ZHENG B Y, GAO X W, PAN R. Sucker rod pump working state diagnosis using motor data and hidden conditional random fields[J]. IEEE Transactions on Industrial Electronics, 2020, 67(9): 7919 – 7928.

[7] ZHENG B Y, GAO X W, LI X Y. Fault detection for sucker rod pump based on motor power[J]. Control Engineering Practice, 2019, 86: 37 – 47.

[8] LI G, JUNG J J. Deep learning for anomaly detection in multivariate time series: approaches, applications, and challenges [J]. Information Fusion, 2023, 91: 93 – 102.

[9] CHEN H S, LIU H, CHU X N, et al. Anomaly detection and critical SCADA parameters identification for wind turbines based on LSTM-AE neural network[J]. Renewable Energy, 2021, 172: 829 – 840.

[10] KILIAN V, PENG Z X, CHRISTOPHER J, et al. Vibration-based anomaly detection using LSTM/SVM approaches[J]. Mechanical Systems and Signal Processing, 2022, 169: 108752.

[11] EL AMOURI H, LAMPERT T, GANÇARSKI P, et al. Constrained DTW preserving shapelets for explainable time-series clustering[J]. Pattern Recognition, 2023, 143: 109804.

[12] YE L X, KEOGH E. Time series shapelets: a novel technique that allows accurate, interpretable and fast classification[J]. Data Mining and Knowledge Discovery, 2011, 22(1/2): 149 – 182.

[13] YAMAGUCHI A, UENO K. Learning time-series shapelets via supervised feature selection[C]//Proceedings of the 2021 SIAM International Conference on Data Mining (SDM). Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2021: 262 – 270.

[14] BEGGEL L, KAUSLER B X, SCHIEGG M, et al. Time series anomaly detection based on shapelet learning[J]. Computational Statistics, 2019, 34(3): 945 – 976.

[15] ZHAO H S, WANG Y L, ZHAN Y, et al. Practical model for energy consumption analysis of beam pumping motor systems and its energy-saving applications[J]. IEEE Transactions on Industry Applications, 2018, 54(2): 1006 – 1016.

[16] HILLS J, LINES J, BARANAUSKAS E, et al. Classification of time series by shapelet transformation[J]. Data Mining and Knowledge Discovery, 2014, 28(4): 851 – 881.

[17] WANG Y, YIN W T, ZENG J S. Global convergence of ADMM in nonconvex nonsmooth optimization[J]. Journal of Scientific Computing, 2019, 78(1): 29 – 63.

(责任编辑:宋颖韬)