DOI:10.13196/j.cims.2013.12.yaoxinxin.2993.9.2013129

面向设备群体的工况数据异常检测方法

姚欣歆，刘英博，赵　炯，胡游乐，张　力
（清华大学 软件学院，北京　100084）

摘　要：为了全面便捷地检测出有效的潜在的设备异常，基于设备大量有价值的工况数据，对设备异常检测方法进行研究，提出一种面向设备群体的工况数据异常检测方法。通过对设备状态监测数据潜在异常特征进行分析挖掘，识别设备运行中的异常状态和用户异常行为，提高产品服务质量。给出了设备异常检测原理，对数据预处理、特征提取、正常特征空间分布建模、异常检测和人工验证过程进行了详细阐述。通过对 5 000 多万条真实监测数据进行异常检测实验，验证了所提方法的可行性和有效性。

关键词：设备群体；工况数据；特征空间；分布建模；异常检测
中图分类号：TP391；TP319　　**文献标志码**：A

Device group-oriented method for abnormal floor data detecting

YAO Xin-xin, LIU Ying-bo, ZHAO Jiong, HU You-le, ZHANG Li
(School of Software, Tsinghua University, Beijing 100084, China)

Abstract: To detect effective potential equipment abnormal comprehensively, the detecting method based on amount of valuable equipment condition monitoring data was researched, and a device group-oriented method for abnormal floor data detecting was proposed. Through analyzing and mining the potential abnormal features of equipment condition monitoring data, the abnormal state and the user abnormal behavior in equipment operation were identified, and the service quality of products was improved. The principle of abnormal detecting was presented, and the processes of data preprocessing, feature extraction, spatial distribution of normal features modeling, abnormal detecting and artificial validation were elaborated in detail. More than 50 million real monitoring data were detected in anomaly detecting experiment, and the feasibility and effectiveness of the method were validated.

Key words: device group; floor data; feature space; distribution modeling; abnormal detecting

0　引言

随着全球经济的快速发展，产品设计制造商和产品用户更加关注产品的维修保养服务，如何减少大型复杂装备的维修费用、缩短服务响应时间和提高设计生产质量是产品生命周期状态监测与运维服务关注的重点和难点。传统的应对方案主要基于设备自身机理，由领域专家为设备关键部件建立针对性的故障检测模型，以发现、诊断并定位故障。

但是，对于大型复杂设备，仅根据产品失效机理分析故障发生现象、提取故障模式非常困难。由于在设计时给出的复杂装备故障失效模式有限，在不同的安装和使用条件下，设备故障模式也会相差很大[1]，而且构建故障模式费时费力，针对每种故障进行建模是不可行的。因此，采用一般的检测方法难以获取设备的所有故障信息。此外，随着设备更新换代越来越快，设备越来越复杂，以往的故障检测模式也不一定适用于新的更复杂的设备。因此，基于模型的故障检测方式在成本、时间、模型覆盖面和持久度方面都有一定的局限性。

收稿日期：2013-07-01；**修订日期**：2013-11-19。Received 01 July 2013; accepted 19 Nov. 2013.
基金项目：国家 863 计划资助项目(2012AA040911, 2012BAF12B23)。**Foundation items**: Project supported by the National High-Tech. R&D Program, China (No. 2012AA040911, 2012BAF12B23).

另一方面，通过挖掘设备正常工作模式发现设备异常现象，可以避开获取故障模式样本的难点，大大减少传统故障建模的时间和成本，在可操作性和应用范围上也有明显的优势。同时，随着物联网和传感器技术的日益成熟，设备状态监测系统在电力、交通运输和机械等众多行业中得到大量应用，为异常检测提供大量设备状态监测数据，作为数据挖掘的基础。此外，与故障检测相比，异常检测虽然不能诊断设备故障，但是可以从中挖掘设备的潜在故障，以弥补故障检测过程中遗漏的故障漏洞。异常检测还可与状态检修技术相结合，实现设备的预测维护和性能退化评估等[2-3]，对降低企业维修成本、提高产品服务质量等具有重大意义。因此，状态监测系统的逐渐成熟，以及对提升产品维修维护和服务质量的迫切需求，都将使异常检测成为今后研究的主流[1]。

目前，在设备异常检测相关研究方面已有一些成熟的技术和成果，主要分为阈值异常检测和数据驱动的异常检测。阈值异常检测是最基础和广泛使用的技术，基于此，衍生了大量相关方法和模型的研究。张宏利等[4]使用可变阈值信息检测器检测设备异常，倪景峰等[5]通过最小二乘支持向量机法计算预测误差是否大于设定阈值，判定数据中是否有异常。阈值异常检测操作简单，但由于设定阈值，需要大量领域知识，检测范围小。挖掘数据内在信息的基于数据驱动的异常检测因适用于大量数据异常检测，无需过多领域知识而逐渐兴起。Wang 等[6]利用正常和异常状态数据的互关联相似度熵构建不同模型以识别异常，该方法需要先识别一定历史数据中的异常数据作为构建模型的基础。Gao 等[7]将航天器相关参数信息聚类为不同行为以挖掘异常，该方法在参数选择上需要了解一定的领域知识，而且参数选择的好坏直接影响异常挖掘的结果。本文所采用的方法属于数据驱动异常检测，不同于上述两种方法，本文方法无需先识别一定历史数据中的异常数据和了解相关领域知识作为异常检测基础，而是采用面向设备群体历史数据而非设备个体历史数据进行检测。这种方法的可操作性强、数据量更大、可挖掘信息更多、适用范围更广，无论工况数据质量好坏，该方法均适用。对于数据质量较差的工况数据，设备个体工况数据量少，且由于设备所处环境不同，不同设备的同种工况数据表现有一定差别，将两两设备数据进行比对以检测异常的方法，是存在很大误差的。而由于设备总量多，使得群体数据量大，个体异常数据会淹没在大量共性数据中，从而可构建有效的正常数据模型并进行异常检测。因此，面向设备群体的大规模数据进行异常检测的方法十分适用于质量较差的设备状态监测数据。目前，相关的异常检测研究还较少。

为了识别设备的潜在异常状态和用户异常行为，提高产品服务质量，本文提出一种面向设备群体状态监测数据的潜在异常特征检测方法，其关键在于在大规模的监测数据中，如何构建合理的正常模型，使其能检测出异常数据。其难点是如何获得构成正常模型的合理元素和构建何种模型。该方法中的数据预处理和特征提取步骤解决了第一个难点；正常特征空间分布建模和异常检测步骤解决了第二个难点。下面对异常检测原理和方法中的数据预处理、特征提取、正常特征空间分布建模、异常检测和人工验证过程进行详细阐述，并针对 5 000 多条真实监测数据进行异常检测实验，以验证该方法在工程应用中的可行性和有效性。

1 问题描述与基本定义

1.1 问题描述

(1)工况数据特点

本文所指设备工况数据为设备监测系统实时监测设备状态并回传的数值型数据，是监测数据的一种，它描述了该设备在一定条件下的工作状态。复杂装备的工况数据一般有以下特点：

1)维度高　不同工况数据为分析设备状态提供了不同的视角，即维度，因此维度高指一台设备的工况种类繁多。例如一台五十铃三桥 46 m 泵车有 270 多个工况采集点，采集到的工况数据共同反映了该泵车的整体运行状态。

2)数据量大　由于设备种类和数量多，每种设备的工况种类多，且监测系统按周期回传工况数据，使得一定时间内累积的设备工况数据量大。例如某工程机械制造企业生产的设备在线 8 万台，每日存储工况数据可达 2 亿条记录。

3)数据类型复杂　不同工况数据的表现形式不同，包含布尔量、模拟量等各种类型，如紧急停止、高压启动等为布尔量；发动机转速、分动箱转速等为模拟量。

4)采集周期不一致　采集周期为设备监测系统回传工况数据的周期。由于工况数据特点和用户关注度的不同，不同工况数据设置的采集周期也不同。例如紧急停止工况数据采集周期不固定，当泵车紧

急停止按钮按下时开始采集回传,而分动箱转速工况数据的采集周期则设置为固定的 1 h。此外,由于一些客观原因导致某些工况数据采集稀疏,数据质量较差,例如发动机转速工况数据每时每刻都会变化,但其采集周期为每小时一次,采集时间间隔久,丢失了很多时间序列信息。

(2)基本假设

本文提出的异常检测方法的基本原理是:在具备大量监测数据的条件下,基于设备群体监测数据,发现群体设备的正常监测数据规律,以此为基础检测潜在的异常个体。因此,本研究工作建立在群体设备的监测数据具有可分析性的基础上。

由于一种设备在投入使用前经出厂检验质量合格,正常使用期间,大部分设备的工作状态都是正常的,因而反映其状态的工况数据也都是正常数据,且具有共性特征。因此,本文假定在群体设备工况数据集中,绝大多数设备的个体工况数据都是正常数据,且正常数据在特征空间分布中总是符合某种规律,而异常数据则不符合这种规律。基于以上假设,针对群体工况数据特点的异常检测,需要解决以下两个问题:

1)定义群体正常工况数据分布模式。该问题的挑战在于:如何准确划定正常和异常数据的边界、如何解决误识噪声数据为异常数据的问题、如何解决模型屏蔽效应等问题。其中模型屏蔽效应问题为因检测区间选择不当,造成检测区间内大量数据为正常数据,异常特征不突出,因而无法检测出异常。

2)基于构建的正常工况数据分布模式,找到一个有效的异常检测方法。该问题的关键在于如何找到一个针对工况数据特点且可操作性强的异常检测方法,在检测到的大量异常中挖掘用户最关注的异常和最有可能存在潜在故障的异常。

1.2　基本定义

定义 1　设备工况数据 d。由于设备监测系统间隔采样,采样时间序列 $T=\{t_1,t_2,\cdots,t_n\}$ 内的设备工况数据为一系列离散时间序列点及其对应的数据值,表示为 $d=[(t_1,v_1),(t_2,v_2),\cdots,(t_n,v_n)]$。其中 $v_1,v_2,\cdots,v_n$ 为时刻 $t_1,t_2,\cdots,t_n$ 下监测系统传回的工况数据值。

定义 2　单设备工况数据 $d_{e,w}$。一台设备的工作状态由多个工况共同表示。单设备工况数据为针对一台设备中特定工况的数据,表示为 $d_{e,w}=[(t_1,v_1),(t_2,v_2),\cdots,(t_n,v_n)]e\in E,w\in W$。其中 $E=\{e_1,e_2,\cdots,e_m\}$ 为设备群集合,集合中每一个元素表示一台设备。$W=\{w_1,w_2,\cdots,w_k\}$ 为该设备群的所有工况集合,集合中每一个元素表示一种工况。

定义 3　多设备工况数据集 D_E。设备群中的设备拥有相同的工况种类。多设备工况数据为针对特定工况,其集合表示为 $D_E=\{d_{e1},d_{e2},\cdots,d_{em}\}$,集合中的每一个元素为设备群中的单设备工况数据。例如针对液压油温度工况,$E=\{11BC53136680,11BC53136681\}$,$D_E=\{d_{11BC53136680},d_{11BC53136691}\}$ 表示两台泵车的液压油温度工况数据。

由于工况数据维度高,无相关性的工况数据不存在可比性;工况数据量大,直接检测所有数据,操作成本高。而选取相关性高的多工况进行建模操作复杂,且选取工况过程中需要一定领域知识,则针对单工况进行检测可操作性更强。因此以下所有问题均对设备群体单工况数据进行讨论。

2　异常检测原理

图 1 为异常检测原理图,其中包括图中间的异常检测原理示意部分和图外围的模块涉及的数据流部分。异常检测原理主要包括数据预处理、特征提取、正常特征空间分布建模、异常检测和人工验证五大模块。其中异常检测模块又由异常检测和异常结果处理两个子模块构成。

(1)*数据预处理模块*　为降低正常特征空间建模的误差和工作量,有必要进行数据预处理。其主要工作为处理并选择适合于异常检测的数据,并将数据按检测区间进行分割。该模块输入监测系统采集到的原始设备群工况数据集 D_E,输出检测区间的数据集 $O_{E,\Delta T}$ 到特征提取模块。

(2)*特征提取*　提取检测区间数据集中每一个检测区间数据特征值。该模块输入检测区间数据集 $O_{E,\Delta T}$,输出设备群的特征空间集 C_E,为构建正常特征空间分布打下基础。

(3)*正常特征空间分布建模*　该模块是异常检测核心模块,主要针对设备群的特征空间集特点,选择并采用合适的建模方法构建正常特征空间分布,为异常检测提供基线模型。该模块输入设备群的特征空间集 C_E,输出正常特征空间分布模型到异常检测子模块。

(4)*异常检测*　该模块也是异常检测核心模块,其目的是检测并挖掘最可能存在潜在故障和用户最关注的异常数据,由初步异常检测和异常结果处理子模块构成。初步异常检测子模块选择并采用有效算法,检测不在正常特征空间分布中的异常数据,该

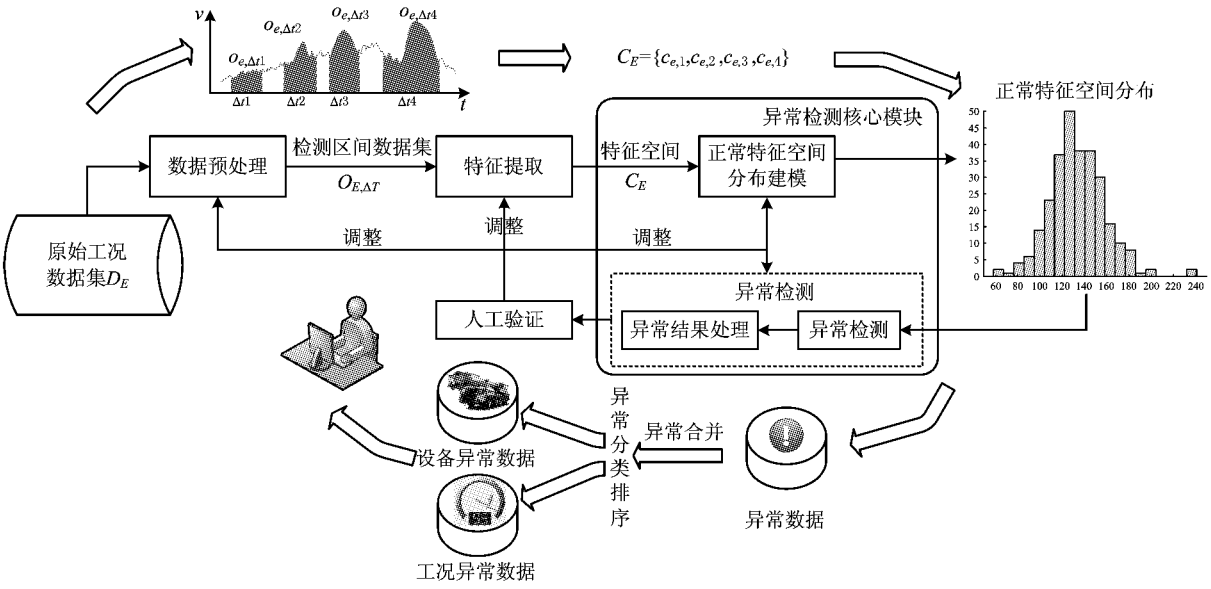


图1 异常检测基本原理

子模块输入特征空间集 C_E 中的特征值和正常特征空间分布模型，输出异常数据集到异常结果处理子模块；异常结果处理子模块主要合并冗余异常数据，对异常进行分类排序以选择有价值的异常数据，减少人工验证工作量。该子模块输入检测到的未经处理的异常数据集，输出统计排序后异常出现次数较多的设备异常数据和工况异常数据到人工验证模块。

(5)人工验证　领域专家通过查看异常检测模块输出的异常数据集，找到异常原因，并对以下可能出现不合理设置的三个模块的相关参数进行反馈调整：数据预处理模块中检测区间的设定、特征提取模块中特征方法的选择和正常特征空间分布建模模块中正常特征空间分布范围的设置，以优化检测结果，减少无用工作，因此该模块是必不可少的一步。

3　异常检测方法

基于上文的异常检测原理给出一个具体的异常检测方法，并对数据预处理、特征提取、正常特征空间分布建模和异常检测四个模块进行详细阐述。

3.1　数据预处理

基于工况数据特点的分析，数据预处理可分为：

(1)处理并选择适合进行异常分析处理的工况数据源

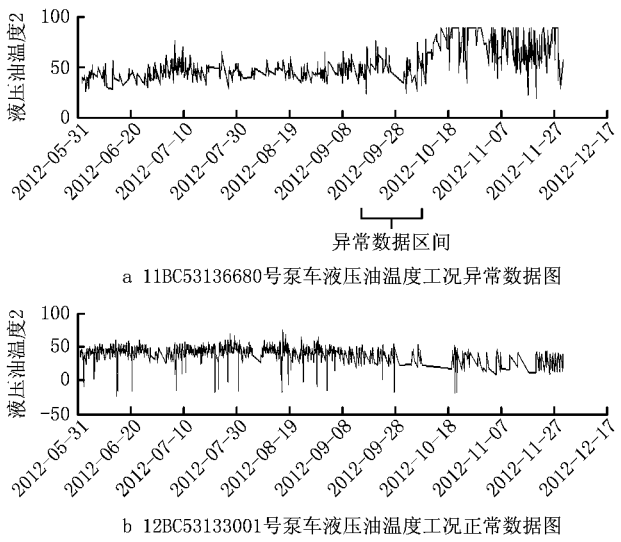
处理并选择好工况数据源是异常检测中十分重要的一步，主要包括数据集成、数据清理、数据变换、数据简化选择工作。其目的在于提高数据质量、减少数据规模、提高检测速度。基于概念树浓缩方法、信息论思想等大量相关方法都可以用于处理相关工作[9-10]，本文不再赘述。

(2)按检测区间分割工况数据

检测区间应为能完整表现所检测异常的最小数据单元。检测区间的长度直接影响异常检测效果。如图 2 所示，假设液压油温度上升数据为异常数据，对应图 2b 中的正常区间数据，图 2a 中所标 9 月 30 日后的区间数据为异常数据。若将检测区间长度设定为整个工作时间，则由于异常数据只占其中一小段，异常数据会被大多数正常数据屏蔽掉，检测不出异常。若长度过小，则不能检测到一段完整的异常。设定合适的检测区间可解决上文所提到的误识噪声问题和模型屏蔽效应问题，为构建正常特征空间分布模型打下良好的基础。对于待检测设备工况数据集，将其分割成合适的检测区间集十分关键。

因此，本文给出以下检测区间分割函数定义，检测区间分割函数的分割原则为将数据集分割成最能体现完整异常特征的区间集，一般由工况异常特征持续时间和异常检测需求等决定，其中工况异常特征持续时间占主导地位。例如图 2a 中所示的异常特征持续了 20 d 左右，分割检测区间长度 20 d 为最佳；当异常检测需求为检测日油耗量是否正常时，检测区间应以 d 为单位。

定义 4　检测区间分割函数 F。检测区间分割函数 F 将原始设备群体工况数据分割成相应的检测区间数据集，表示为：$F(D_E,\Delta_T)\rightarrow O_{E,\Delta T}$。其中分



a 11BC53136680号泵车液压油温度工况异常数据图

b 12BC53133001号泵车液压油温度工况正常数据图

图2　液压油温度工况数据图

割函数 F 有两个输入变量和一个输出变量。F 的两个输入变量为 D_E 和 ΔT。其中:D_E 为上文定义的多设备工况原始数据集,$D_E=\{d_{e1},d_{e2},\cdots,d_{em}\}$;$\Delta T$ 为最优检测区间集,$\Delta T=\{\Delta t_1,\Delta t_2,\cdots,\Delta t_p\}$。两两检测区间无交集,则 $\forall\Delta t_1,\forall\Delta t_2\in\Delta T,\Delta t_1\cap\Delta t_2=\varnothing$。根据分割原则,不同的工况异常特征和检测需求将导致不同长度的检测区间,则不同数据的最优检测区间集 ΔT 也不同。F 的输出变量 $O_{E,\Delta T}$ 为分割后的检测区间数据集,

$$O_{E,\Delta T}=\{o_{e1,\Delta t1},o_{e1,\Delta t2},\cdots,o_{e1,\Delta tp},o_{e2,\Delta t1},o_{e2,\Delta t2},\cdots,o_{e2,\Delta tp},,\cdots,o_{em,\Delta t1},o_{em,\Delta t2},\cdots,o_{em,\Delta tp}\}。$$

集合中的每个元素都为一个检测区间数据。$o_{e1,\Delta t1}$ 表示原始数据 d_{e1} 中 $\Delta t1$ 时间内的所有时间点对集,$o_{e1,\Delta t1}=\{(t_a,v_a),(t_b,v_b),\cdots,(t_c,v_c)\}$,$t_a+\Delta t_1\geqslant t_c$。

图 3 为检测区间分割结果示意图。假设根据设备编号为 11BC53136680 的泵车液压油温度工况异常特点和检测需求,确定检测区间集 $\Delta T=\{\Delta t1$:12/6/20－12/7/26,$\Delta t2$:12/8/18－12/9/4,$\Delta t3$:12/9/30－12/10/20,$\Delta t4$:12/11/5－12/11/30$\}$,使用上述分割函数对其进行分割处理,由于设备群 E 中只有这一台设备,$E=\{e$:11BC53136680$\}$,则检测区间集为 $O_{E,\Delta T}=\{o_{e,\Delta t1},o_{e,\Delta t2},o_{e,\Delta t3},o_{e,\Delta t4}\}$。一般情况下,一个设备的检测区间集 ΔT 中的元素是连续不相交的,此处为便于区分四个检测区间,将其绘制成非连续数据集。

3.2　特征提取

设备工况数据的特征主要有三类:

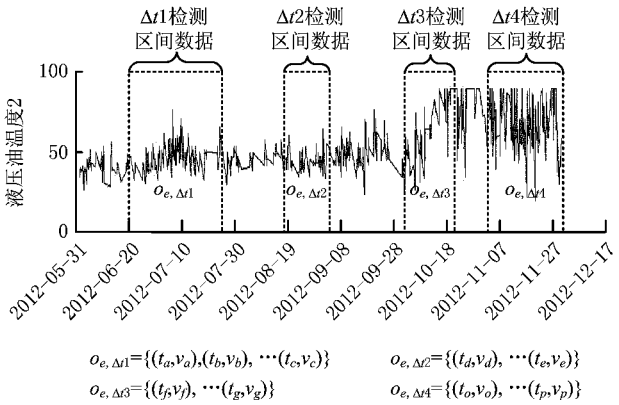


图3　液压油温度检测区间数据集示意图

(1)**工况取值分布特征**　表示检测区间数据的集中程度、分散范围等分布状态,如平均值、方差等。

(2)**工况取值趋势特征**　表示检测区间数据随时间的变化趋向,如单调递增、单调递减等。

(3)**工况采样特征**　表示检测区间数据的采样信息,如采集频率等。

针对不同工况数据类型,可以采用不同的特征提取方法,本文列出了常用的七种方法。

针对每种特征方法,对检测区间数据集进行特征提取,都会生成一个相对应的特征空间集,特征空间集是设备群正常特征空间分布建模的基础。下面给出其定义与求解过程。

定义 5　特征空间集 C_E。设备群特征空间集为采用特征方法集 $FE=\{$Max, Min, Avg, Stdev, Desc, Asc$\}$中的一种特征提取方法,对检测区间数据集 $O_{E,\Delta T}$ 中每个检测区间的数据元素提取特征,所生成的特征点值集即特征空间 C_E,

$$C_E=\{c_{e1,1},c_{e1,2},\cdots,c_{e1,p},c_{e2,1},c_{e2,2},\cdots,c_{e2,p},\cdots,c_{em,1},c_{em,2},\cdots,c_{em,p}\}。$$

其中每个检测区间数据会被映射到特征空间中的一个点上,表现了该检测区间数据的特征。如特征方法选 Avg 方法,则 $c_{e1,1}$ 表示检测区间数据 $o_{e1,\Delta t1}=\{(t_a,v_a),(t_b,v_b),\cdots,(t_c,v_c)\}$ 中所有时刻(n 个时刻)取值的平均值,即 $c_{e1,1}=\dfrac{v_a+v_b+\cdots+v_c}{n}$。对于液压油温度工况数据分割后的检测区间集,经提取平均值特征后,则变为特征空间 $C_E=\{c_{e,1},c_{e,2},c_{e,3},c_{e,4}\}$。

3.3　正常特征空间分布建模

选择并采用基于统计的正态分布模型,以构建设备群正常特征空间分布。

图 4 为某工程机械制造企业监测系统回传的所有泵车的比例阀电流波动示意图。比例阀电流数据反映了泵送速度，其波动情况可由方差特征表示。横轴表示比例阀电流方差值，纵轴为车辆数目。由图 4 可知，比例阀电流波动情况符合正态分布，大量车的比例阀电流波动值集中在 80～180 之间。事实上，比例阀电流波动太大或者波动太小都是非正常现象，异常比例阀电流波动值为该图两边较少的车辆数据。因此，基于统计的正态分布模型十分适用于工业中监测工况数据的异常检测。

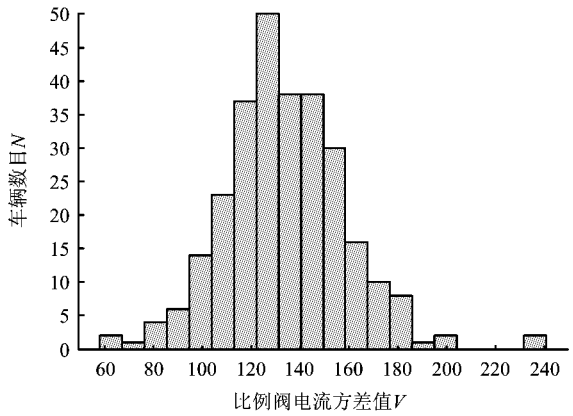


图4 比例阀电流波动情况图

下面给出正常特征空间分布模型具体构建过程：

(1)构建设备群单特征空间正态分布

$$F(x)=\int_{-\infty}^{x}f(x)\mathrm{d}x=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x}e^{\frac{(x-\mu)^2}{2\sigma^2}}\mathrm{d}x。\quad(1)$$

式中：μ 为设备群特征空间 C_E 中所有特征点值的平均值，σ 为其方差。

(2)采用工业中常用的 3σ 原理构建正常特征空间分布

3σ 为基于设备群特征空间符合正态分布的前提，提出正常特征点值 c 落入区间$[\mu-3\sigma,\mu+3\sigma]$的概率为

$$Pr(\mu-3\sigma\leqslant c\leqslant\mu+3\sigma)=0.997\ 3。\quad(2)$$

则正常特征空间分布区间应为$[\mu-3\sigma,\mu+3\sigma]$。$3\sigma$ 原理符合上文所分析的实际正常车辆分布情况，且 3σ 原理在实际计算过程中简单方便、可操作性强，十分适用于监测大数据的异常检测。

3.4 异常检测

异常检测主要分为初步异常检测和异常结果处理两步，其目的在于检测挖掘用户最关注的异常和最有可能存在潜在故障的异常数据，并减少领域专家人工验证的工作量，提高效率。

3.4.1 初步异常检测

根据假设和所构建的正常特征空间分布模型，给出以下初步异常检测方法。

方法输入：待检测特征值 $c_{e,p}$，正常特征空间分布中的平均值 μ 和方差 σ。

方法输出：表示其是否为异常数据的布尔值，1 表示异常，0 表示正常。

则异常检测函数：

$$f(c_{e,p},\mu,\sigma)=\begin{cases}1, & c_{e,p}>\mu+3\sigma \text{ or } c_{e,p}<\mu-3\sigma;\\ 0, & \mu-3\sigma\leqslant c_{e,p}\leqslant\mu+3\sigma。\end{cases}\quad(3)$$

若检测数据结果为异常，则获取其异常检测区间数据，加入到异常数据集中。异常数据集中记录了所有设备不同工况采用不同特征提取方法所检测出的异常检测区间集。

异常检测分为两种情况：

(1)批量检测历史数据异常

对历史数据特征空间集中的每一个特征点值进行异常检测。

(2)增量检测实时数据异常

实时获取监测系统回传的工况数据，并计算其长度是否达到一个完整的检测区间长度，当达到时对其进行分割并提取特征值。由于新增的单个检测数据特征值不会影响整体的正常特征空间分布，可采用上述异常检测方法对增量特征点值进行增量地异常检测。

3.4.2 异常结果处理

(1)合并冗余异常

采用不同的特征提取方法和不合理的检测区间会造成大量冗余的异常结果，浪费大量的人工验证时间和工作，因此需要合并冗余异常结果。

合并分为三种情况，如图 5 所示。

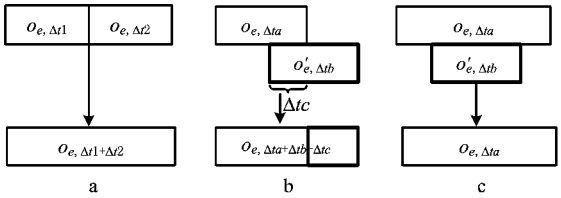


图5 合并异常结果的三种情况

1)特定特征异常检测过程中，当同一设备异常检测区间数据集中出现相邻检测区间数据时，将相邻的两个或多个异常检测区间数据合并为一个异常数据。如图 5a 所示，在平均值特征异常检测中，

$o_{e,\Delta t1}$ 和 $o_{e,\Delta t2}$ 为相邻的异常检测区间数据,则将其合并为 $o_{e,\Delta t1+\Delta t2}$。

2)不同特征的异常检测过程中,当同一设备多个异常检测区间数据集中存在交集关系时,将相交检测区间数据的并集作为一个异常数据。如图 5b 所示,在平均值特征异常检测中,$o_{e,\Delta ta}$ 为一个异常检测区间数据;在方差特征异常检测中,$o'_{e,\Delta tb}$ 为一个异常检测区间数据,且 $o_{e,\Delta ta} \cap O, o'_{e,\Delta tb} \neq \varnothing$,$\Delta ta \cap \Delta tb = \Delta tc$,则将其合并为 $o_{e,\Delta ta+\Delta tb-\Delta tc}$。

3)不同特征的异常检测过程中,当同一设备多个异常检测区间数据集中存在真子集关系时,保留超集作为一个异常数据。如图 5c 所示,$o_{e,\Delta tb} \subseteq o_{e,\Delta ta}$,则将其合并为 $o_{e,\Delta ta}$。

(2)异常统计分类

异常数量较多的工况说明与其关联的一个或多个构件易磨损,是产生潜在故障的重大隐患工况。而异常数量较多的设备可分为两种情况:①设备某工况频繁出现异常,它可能反映了因磨损、疲劳等带来的潜在故障或用户频繁的不正当操作行为;②设备多种工况出现异常,它可能反映了由多工况共同影响的潜在故障。以上分析对于潜在故障的挖掘十分重要,因此异常检测的最后一步为将检测和合并后的异常数据按设备、工况和异常类型进行统计排序和分类展示,将出现异常数量较多的工况和设备异常数据提供给相关领域专家进行人工验证。

4　系统实现和实验结果

4.1　实验系统

基于上述异常检测原理,实现了用于异常检测的数据分析系统。该系统采用 JavaEE 技术,用 LaUD 数据库存储工况大数据,用 Oracle 数据库存储异常检测结果等。实验数据为某工程机械制造企业 279 台泵车 6 个月的 200 多种工况数据,共 5 283 万个设备工况数据。每个工况数据都记录了设备工况单时刻的取值。实验过程如下:①数据预处理过程,对工况数据进行简单的分类清理工作。首先将工况数据按不同数据类型(布尔量、模拟量、固有量等)和所属部件等分类方式进行分类,为不同类型数据特征提取打好基础。然后剔除数值长期不变等无用数据和过高过低的噪声数据等,减少数据分析工作量。此外,由于泵车工作时间一般以月为单位,且实验数据采集稀疏,日采集数据量较少,无法体现异常数据的完整特征。最终经统计计算,将检测区间选定为一个月,以进行数据分割。②特征提取过程,系统起初采用表 1 所示的 7 种特征提取方法,但一些方法效果不佳,经人工验证调整后,最终选择了其中最有效的 3 种方法——方差、升高率和采集频率。分别计算其正常特征空间分布并初步检测异常。系统共检测出4 000多种异常数据,分别属于 278 辆泵车,存储在异常数据表中。随后,系统将检测出的异常进行合并,并将合并后的异常数据按车辆和工况进行了统计排序,生成设备异常数据表和工况异常数据表,以查看不同情况下异常出现较多的工况和车辆,其中 33 辆车、59 种工况出现的异常较多。为了方便领域专家分析异常,还实现了异常可视化工具,通过将异常数据和随机抽取的正常数据进行对比,更直观地展示了异常数据。图 6 为系统界面图,左边为功能选择区域,右边为结果展示区域,当前图的右边区域展示了设备异常数据表中异常最多的前 29 辆车的异常信息。

表 1　七种特征提取方法

名称	介绍	适用数据类型	检测意义
最大值 Max	数据中最大可达值	模拟量	检测数值过高异常。如检测发动机转速过高情况等
最小值 Min	数据中最小可达值	模拟量	检测数值过低异常。如检测电流过小情况等
平均值 Avg	所有数据的平均值	布尔量和模拟量	检测数据集中程度异常。如检测日平均耗油量等
方差 Stdev	所有数据的方差值	布尔量和模拟量	检测数据变化程度异常。如检测倾角变化幅度异常等
下降率 Desc	后一个数据小于前一个数据的统计概率	模拟量	检测数据跳变程度和单调性异常

续表 1

升高率 Asc	后一个数据大于前一个数据的统计概率	模拟量	检测数据跳变程度和单调性异常。如检测工作时长等
采集频率 Freq	每秒采样个数	所有类型	检测采集频率和其潜在异常现象，如检测频繁报警情况、传感器失常等

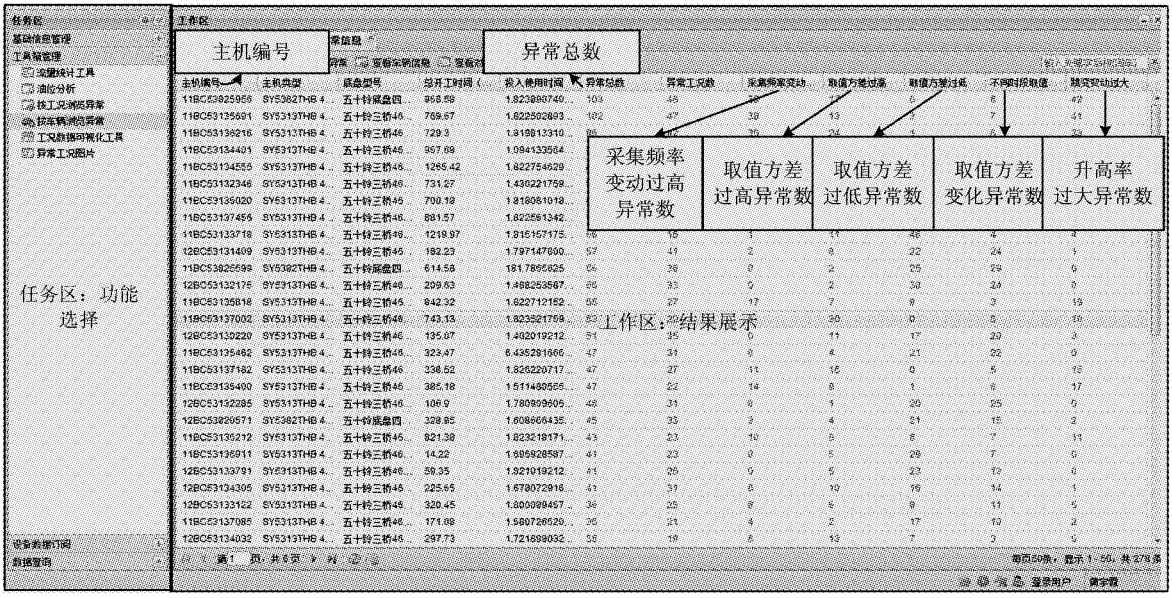


图6 系统界面图

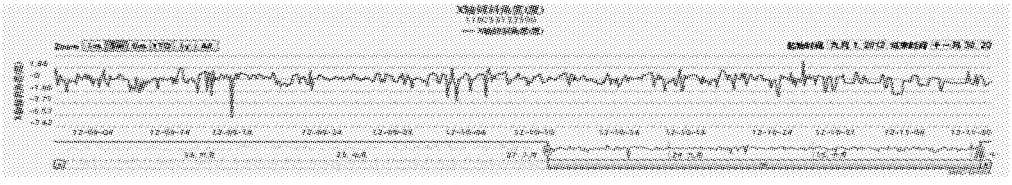
4.2 实验结果分析

经领域专家的分析和实验检测出的异常数据，对识别用户异常行为、发现安全隐患和促进质量改进等方面十分有帮助，从而证明了所检测的异常数据的价值和异常检测方法的有效性。

例如通过取值方差过小异常检测，本文发现 x 轴倾斜角度工况出现异常情况。经异常检测方法所得的该工况正常取值方差特征空间范围应在 0.3～0.5 之间，因此，如图 7b 中的正常工况数据所示：编号为 11BC531133590 泵车的 x 轴倾斜角度值会在 0 值范围内小幅度波动。但图 7a 中编号为 11BC531132319 泵车的 x 轴倾斜角度一直为 0，是异常数据。

检测出的异常数据经由领域专家分析验证表明，该工况数据为反映泵车离地倾斜角度而设计，其正常的表现形式确实与分析所得结果一致，其值不

a x轴倾斜角度工况异常数据



b x轴倾斜角度工况正常数据

图7 x轴倾斜角度无变化异常图

断波动的原因是当泵车倾斜角度超过一定值时,泵车会自动进行调整,保持平衡。经调查发现,异常现象是由于客户希望泵车长期运行,自行拔出倾角传感器造成的,属于不正当异常操作行为。这种行为会使得泵车防倾覆功能失灵,产生巨大的安全隐患,若没有及时发现,则后果不堪设想。

在促进质量改进方面,通过升高率过小异常检测,发现主机工作时间异常情况。经人工观察验证,检测出的异常数据有很多跳变,确实与主机工作时间正常工况数据的单调递增特征相悖。经了解,确定其原因为泵车错误回传压路机信源,目前公司正在解决该问题。

以上示例充分表明了异常检测的重要价值和本文所提出的异常检测方法的可行性和有效性。

5　结束语

本文主要针对监测工况数据,从数据预处理、特征提取、正常特征空间分布建模、异常检测和人工验证反馈五大过程论述了一种面向群体设备的工况数据异常检测方法。通过对真实数据进行异常检测实验,验证了该方法的可行性和有效性。有别于传统的故障检测模型,该方法在可操作性、覆盖面和持续性上有很大优势;且可检测设备潜在异常,识别用户行为,发现安全隐患和促进质量改进,有助于企业提高服务质量。

今后的重点研究内容包括根据工况数据异常特征表现,动态确定最优异常检测区间长度,进行多工况关联异常检测以及将异常检测过程与状态检修相关技术和模型相结合等。

参考文献:

[1] MIN Rui, XU Liming, WEI Jujun, et al. Anomaly detection of running state of machining system[J]. Chinese Journal of Scientific Instrument, 2006,27(6):1777-1778 (in Chinese). [闵　睿,许黎明,魏巨隽,等. 机械加工系统运行状态异常检测方法的研究[J]. 仪器仪表学报,2006,27(6):1777-1778.]

[2] XIAO Gui. Anlmaly detection methods and measures for operating state of mechanical processing system[J]. Technological Development of Enterprise, 2011, 30(3): 83-84 (in Chinese). [肖　贵. 机械加工系统运行状态异常检测方法及措施探讨[J]. 企业技术开发, 2011,30(3):83-84.]

[3] MCARTHUR S D J, BOOTH C D, MCDONALD J R, et al. An Agent-based anomaly detection architecture for condition monitoring[J]. IEEE Transactions on Power Systems, 2005, 20(4):1675-1682.

[4] ZHANG Hongli, LIU Shulin, JIAO Wenhui, et al. Equipment abnormal degree detection approach based on variable threshold information detector[J]. Journal of Mechanical Engineering, 2013,49(8): 25-30 (in Chinese). [张宏利,刘树林,缴文会,等. 基于可变阈值信息检测器的设备异常度检测方法[J]. 机械工程学报,2013,49(8):25-30.]

[5] NI Jingfeng, LIU Lihua, GU Yujiong. Outliers detection in time series of measured data based on least square support vector machine algorithm[J]. Journal of North China Electric Power University, 2008,35(3): 62-66 (in Chinese). [倪景峰,刘丽华,顾煜炯. 基于最小二乘支持向量机算法的测量数据时序异常检测方法[J]. 华北电力大学学报,2008,35(3):62-66.]

[6] WANG Tianyang, CHENG Weidong, LI Jianyong, et al. Anomaly detection for equipment condition via cross-correlation approximate entropy[C]//Proceedings of 2011 International Conference on Management Science and Industrial Engineering. Washington, D. C., USA: IEEE, 2011: 52-55.

[7] GAO Yu, YANG Tianshe, XU Minqiang, et al. An unsupervised anomaly detection approach for spacecraft based on normal behavior clustering[C]//Proceedings of the 5th International Conference on Intelligent Computation Technology and Automation. Washington, D. C., USA: IEEE, 2012: 478-481.

[8] FEI Heliang, XU Xiaoling, LU Xiangwei. Masking effect on tests for outliers[J]. Chinese Journal of Applied Probability and Statistics, 2002,18(2): 141-151 (in Chinese). [费鹤良,徐晓岭,陆向薇. 异常数据检验的屏蔽效应[J]. 应用概率统计, 2002,18(2):141-151.]

[9] LIU Mingji, WANG Xiufeng, HUANG Yalou. Data preprocessing in data mining[J]. Computer Science, 2000,27(4): 54-57 (in Chinese). [刘明吉,王秀蜂,黄亚楼. 数据挖掘中的数据预处理[J]. 计算机科学,2000,27(4):54-57.]

[10] WANG Daling, YU Ge, BAO Yubin, et al. A representation about domain knowledge for reprocesses of data mining[J]. Mini-Micro Systems, 2003,24(5): 863-868 (in Chinese). [王大玲,于　戈,鲍玉斌,等. 一种面向数据挖掘预处理过程的领域知识的分类及表示[J]. 小型微型计算机系统,2003,24(5): 863-868.]

作者简介:

姚欣歆(1990—),女,北京人,硕士研究生,研究方向:工作流技术、制造业信息化系统,E-mail:yxx_happylife2010@163.com;

刘英博(1978—),男,湖南长沙人,助理研究员,研究方向:业务过程管理、制造业信息化、软件工程;

赵　炯(1988—),男,陕西宝鸡人,硕士研究生,研究方向:产品全生命周期管理、制造业信息化系统;

胡游乐(1989—),男,福建古田人,硕士研究生,研究方向:产品全生命周期管理、制造业信息化系统;

张　力(1960—),男,安徽绩溪人,副教授,硕士生导师,研究方向:产品全生命周期管理、工作流、维修服务系统。