文章编号:1671-4598(2023)02-0041-07 DOI:10.16526/j.cnki.11—4762/tp.2023.02.007 中图分类号:TP312 文献标识码:A

基于滑动窗口的直升机序列异常检测算法

赵子晗，段同乐，张冬宁
（中国电子科技集团公司 第54研究所，石家庄 050081）

摘要：无标签的序列在异常检测算法中往往存在着对数据的信息掌握不全面、不能合理使用的情况，而采用深度学习的技术实现检测时往往对其计算的解释性欠佳；对于攻克这些难题，以直升机飞行数据为例对时间序列的异常检测问题展开了深入研究，并利用Iforest算法和PCA算法，给出了一个采用滑动窗口的时间序列异常检测方法，利用从滑动窗口采集信息的时间变化状态等数据信息，将序列异常检测问题转换为点异常检测问题；同时以auc评分为衡量标准，从带有时刻特殊标签的多个数据集上检验了检测效率的提高；在无标签的直升机飞行数据集上进行实验，验证了算法的有效性，并通过对比检测过程中不同特征变量的变化情况，从算法层面和现实层面上阐述了算法的可解释性。

关键词：异常检测；序列异常；点异常；直升机飞行数据

Helicopter Sequence Anomaly Detection Algorithm Based on Sliding Window

ZHAO Zihan，DUAN Tongle，ZHANG Dongning
（The 54th Research Institute of China Electronics Technology Corporation，Shijiazhuang 050081，China）

Abstract：In the unlabeled sequence anomaly detection problem，the acquisition of data sequence features is not sufficient and cannot be effectively used，and the deep learning method is often used to detect the algorithm has poor interpretation. In order to solve the above problems，the helicopter flight data was taken as an example to study the anomaly detection of time series. Based on Iforest algorithm and PCA algorithm，a sliding window based sequence anomaly detection algorithm was proposed. By extracting the fluctuation and statistical information of the data through the sliding window，the sequence anomaly detection problem was transformed into a point anomaly detection problem. auc score was selected as the measurement standard to verify the improvement of the detection efficiency of the algorithm on multiple data sets with abnormal labels. Experiments were carried out on the unlabeled helicopter flight data set to verify the effectiveness of the algorithm. By comparing the changes of different characteristic variables in the detection process，the interpretability of the algorithm was illustrated from the algorithm level and the practical level.

Keywords：anomaly detection；abnormal sequence；anomalies；helicopter flight data

0 引言

随着科学技术的发展和人类社会的进步，我们不管是在科技领域还是生活领域都积累了大量的数据信息，而且数据的规模越来越大，在如今数据爆发式增长的情况下，如何管理好并应用好这些数据就显得尤为重要。在这个需求基础上，数据分析等相关技术应运而生[1-2]。大数据分析的广泛应用与进展，也导致了科研人员们针对于军事工程应用中海量数据的管理问题有了全新的认识与要求，而大数据挖掘中的一项十分关键的分支应用便是异常检测，异常检测技术对于机械故障诊断、疾病监测、保险欺骗检测以及身份辨别等领域都发挥着相当重要的作用[3-6]，对于军事应用中的海量数据而言，异常数据中往往蕴含着显著的行为信息，如何提取合适的异常特征并针对无标签数据进行有效率的异常检测，成为了当前面临的难题[7-8]。

当前无标签的序列在异常检测算法中往往存在着对数据的信息掌握不全面、不能合理使用的情况，采用深度学习方法又面临着算法可解释性差等问题[10-14]，基于此我们以直升机飞行数据为例对序列异常检测进行研究，采用基于滑动窗口的直升机序列异常检测算法，提升算法检测效率，实现算法优化；同时选取特征变量，通过对比分析阐述算法的有效性和可解释性[15-18]。

1 飞行数据异常检测技术

1.1 异常及异常检测相关定义

在异常检测技术中，异常是指信息中不满足所规定的正常行为的状态，在一般过程中，信息通常是由一个或多个常规的形成机制产生的，其他的形成机制所产生的信息，一般可视为异常数据。所以，当某些数据点明确的区别或者脱离了通常的点集时，我们就可以大胆猜测其为异常模式所产生的。而序列性异常则是指在一定时刻上，或是在相对空间上具有天然序列性特征的数据。这些数列既可以是单特征变量的，也可以是多特征变量的[19-22]。系列中异常数据产生的因素也有许多，其中主要包括以下原因：

因为工作的疏忽，造成信息的阅读、录入、统计等产生的错误；因为不同的数据库系统的度量内容和时间通常

收稿日期:2022-11-16； 修回日期:2022-12-12。

作者简介:赵子晗(1997-)，男，河北石家庄人，硕士研究生，主要从事大数据分析方向的研究。

引用格式:赵子晗，段同乐，张冬宁.基于滑动窗口的直升机序列异常检测算法[J].计算机测量与控制，2023，31(2)：41-47，54.

并不相同，可能导致了在合并来自不同数据库系统的信息上出现的问题；因为其内部结构的许多内部特征，如上下文关系、因果关系等难以避免的序列特征特异性产生的错误。

而现如今针对飞行序列异常检测通常要面临以下两个主要的难题：

首先，不同于测试用的有标签的明确数据集，应用于工业工程生产中的数据集往往同时具备数据量非常庞大和缺乏标签这两个特征。以直升机的飞行数据收集工作为例，数据量往往可以到达百万量级，但因为给数据打标签的工作通常是由行业内相关专家针对相应的特征变量手动进行，所以要得到一个具有准确异常标签的训练数据集，往往需要花费非常大的时间代价。由于异常的形成因素很多，所以收集已打好标签并横向上涵盖该时间切片情况下，任何可能的异常行为数据往往比收集带标记的正常数据的困难更大，但总的来说，针对在大数据环境下的无标记数据，相关标签的稀缺性使得异常检测的难度骤增[23]。

其次是数据类型的动态变化特性。在很多问题上我们都无法单纯地从数据模型展开解释，需要从其生成的时间流程、行为等来确定异常，而加入了时间的概念后，现阶段识别或标签出来的异常的行为也不一定在下一个时间结点下依然存在着意义，也因此在纵向上统一特征对应的异常界定也很困难，正常数据和异常数据边界的不确定性会导致随着动态数据的识别训练过程中不断增加新类型的异常模式。

1.2 国内外研究现状

随着科技的发展，国内外对于飞行数据的异常检测技术都有了长足的发展：国外对直升机状态的检测从最初依赖工人专业素养和工作经验的的“看”“嗅”“听”“摸”的人工模式逐渐发展为利用传感器与计算机设备相结合的HUMS技术，对于直升机实时监测分析的能力得到显著提高。在这期间 Guanguli 和 Chopra 等人建立了非线性气动弹性方程，模拟了质量块丢失、桨叶吸潮、变距拉杆损坏等具体的故障[24]；B. V. Jammu 提出里 SBCN 神经网络用于 OH-58A 直升机的诊断；Mao Yang 和 Chopra 等在直升机上对旋翼和机身耦合进行了异常检测等[25]。而我国在 Hums 方面的研究开展的比较晚，直升机故障检测技术也发展的相对缓慢。“小样本、贫信息”的灰色系统理论逐渐应用于直升机领域[26]；姚飞虎在盲源分离的人工免疫技术的基础上创建的旋翼故障诊断方法[27]；邓升平在模拟旋翼不平衡实验中建立的支持向量机和广义神经网络的两种故障诊断模型[28-29]等，都为相关领域提供了坚实的基础。

因飞行数据异常检测中异常标签的匮乏，监督型学习方法不能很好地发挥其算法优势，通常要使用已知的正常样本数据来进行学习检测，当前在点异常检测领域所采用的方法基本可分为三类，即基于密度或超平面划分的方法、基于线性模型的方法和基于在线计算的方法：

首先是基于密度或超平面划分的方法中比较有代表性的三个算法：LOF 算法、KNN 算法和 Iforest 算法。其中 LOF 方法（Local Outlier Factor 局部离群因子检测）是根据密度的离群点测量技术中一个常用方法。其算法主要好处在于：它同时兼顾了数据子集的局部与全局特征。LOF 由于性能好，因此特别适合于中高维的数据子集；KNN 算法（k-NearestNeighbor 算法）又称 k-近邻算法。其算法原理是对信息的排序。使用 KNN 算法测试时序数据异常值的优点在于训练时间较短，对数值无假设，准确率高。比较适于对样本容量较大的雷雨自动分析，也可以进行非线性回归，但缺点是运算工作量较大，对稀有类别的数据精确度低，可解释性也较差；Iforest 方法是一种基于集成学习技术的快速异常分析方法，既不需要数学模型又不需要有标签的训练，同时具有线性的时间复杂度和高准确性。但是 Iforest 不能使用太大维度的数据。因为每次切数据都是随机选择某个维度，建了树之后依然有大量的维度数据不能被利用。并且高维数据还可能存在大量噪音维度或者无关维度，使得树的构建难度增大。Iforest 算法的确在异常检验领域中发挥了很大影响，促进了重心推断理论的进展，而且在分类聚类和异常检验领域中都有了明显的成效。

基于线性模型的代表性算法之一为 PCA（principal component analysis）即主成分分析方法，是目前最为广泛应用的数据降维技术。PCA 是一个基于目标数据特征性的最佳正交变换，称它为最佳正交变换主要因为它具备以下较好的特点：转换后与新的能量正交或不有关；转换矢量更趋平衡、能量更趋集中等。PCA 由于简单而有效，广泛应用于数据处理中特征选取、数据压缩等各个方面。

基于在线计算的方法如 LODA，除了快速和准确的特性之外，LODA 还能够对丢失变量的数据操作和更新。此外，LODA 可以识别出被仔细检查的样本与大多数样本不同的特征。当目标是找出导致异常的原因时，此功能非常有用。

上述方法主要是以将点异常分析的研究为重点目标，而在针对无标签数据的序列异常分析方面，现阶段使用的主要是神经网络的方法，包括了自编码网络系统、对抗网络系统和循环神经网络等最先进的深入学习模式，以变分自编码网络系统为例，它融合了机器学习与贝叶斯学习二者的优势，对于异常情况的模拟训练有着更好的拟合效果，且可以充分发挥贝叶斯方法针对小样本学习的稳定性。再比如，训练神经网络可以广泛应用于处理各种序列数据现象的神经网络框架中，但一直无法解决传统循环神经网络在训练过程中所存在的时间梯度消失的问题。总的来说，现阶段用神经网络的方式进行飞行数据序列异常检测是一种非常值得研究应用的发展方向。

2 算法设计与实现

序列异常检测中数据有一定顺序特征，导致了传统的异常检测技术单一地解析各种数据实例，却忽视了数据的

序列特性，从而导致测试的复杂度和准确率都差强人意。而常用的神经网络的方法虽然一定意义上解决了序列特性，但是受限于算法本身，其可解释性较差，对于工业级数据并不具备很好的普适性。

基于此背景下，我们提出并证明了一种基于滑动窗口的序列异常检测算法，通过分别引入两个滑动窗口来满足对于异常检测所需特征的提取和多提取特征的时间关联性需求，滑动窗口的引入获取了序列统计特征，充分提取了时间片段的统计特征，从而实现了序列异常检测问题到点异常检测问题的转换，在点异常检测问题的基础上我们就能有更多的验证手段和检测方法来验证检测效率的提升。我们在 annthyroid、arrhythmia、breastw、cardio、mammography、musk、pendigits、pima、satellite、satimage-2、seismic _ bumps、shuttle、thyroid、wbc 这 14 个带标签的公开数据集上，按照各个数据集已知的异常比例对各个算法设置异常比例参数为运行算法计算 auc 评分并取均值，并统计每个算法能在多少数据集上获得最高评分。其结果见表 1。

表 1　算法选择标准

异常检测算法	平均 auc 评分	获得最高 auc 评分的数据集个数
LOF	0.607 9	0
PCA	0.853 7	3
KNN	0.758 3	2
IFOREST	0.880 3	7
LODA	0.758 4	2

通过综合比对算法检测的评分和算法能够获得最高评分的数据集个数后，我们最终在算法内部选择使用 Iforest 和 PCA 与滑动提取序列特征的窗口相结合。

所采用的序列异常检测算法流程如图 1 所示。

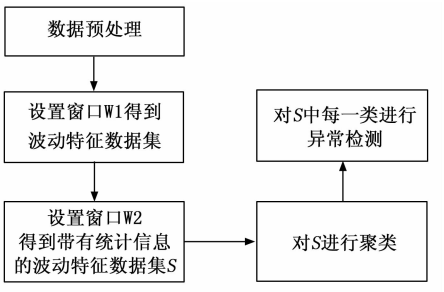

图 1　序列异常检测算法流程图

1）数据预处理。将原始数据按照序列顺序进行筛选，即将数据划分为某个平台仿真实现的某时间区间内的全部数据，并标注相应的时间关系。

2）特征提取。通过提取数据的波动情况和数据的统计信息来完成特征提取的任务。通过计算时间窗口内的 2-范数值和范数变化率，得到数据的波动情况。

$$\|\boldsymbol{X}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \tag{1}$$

其中：$X=\{x_1, x_2, \cdots, x_n\}$。范数变化率即为相邻时间窗口间的范数值之差。

再使用时间窗口获取数据统计信息。数据的统计信息可以反映该段时间内数据的总体状况。本算法中选取的特征值见表 2。

表 2　统计特征表

特征值名称	反映的数据特征
均值	数据的集中趋势
标准差	度量样本点与期望值的偏差
极差	统计样本点的变异数量
最大值、最小值	衡量样本点的域值

用 $T=\{T_1, T_2, \cdots, T_{n-1}, T_n\}$ 来表示 m 维时间序列，T_i 代表第 i 时刻参数项数据值。其中：

$$T_i = \{t_i^1, t_i^2, \cdots, t_i^{m-1}, t_i^m\} \tag{2}$$

t_i^j 为数据的第 i 时刻，第 j 项参数的值。

引入滑动窗口 W_1，设置窗口大小为 k、滑动步长为 step _ 1，计算每 k 个时间步内，第 j 项的参数数据的范数值和范数变化率。

$$N^j(T_k) = \sqrt{(t_i^j)^2 + (t_{i+1}^j)^2 + \cdots + (t_{i+k-1}^j)^2} \tag{3}$$

$$DN^j(T_k) = N^j(T_k) - N^j(T_{k-1}) \tag{4}$$

通过公式（3）、（4）计算得到波动特征数据集 D，D 的维度为 $u * v$，其中 m 为选取飞行特征数量，

$$u = \frac{n - W_1}{\text{step}\_1} + 1 \tag{5}$$

$$v = 2 * m \tag{6}$$

n 为总数据个数。

引入第二个滑动窗口 W_2，设置窗口大小为 k^*、滑动步长为 step _ 2，计算波动特征数据集中每 k^* 个时间步内，各项参数数据的统计特征值，最终得到带有波动统计信息的特征数据集 S。

S 的维度为 $x * y$，其中：

$$x = \frac{(u - W_2)}{\text{step}\_2} + 1 \tag{7}$$

$$y = v * z \tag{8}$$

z 为所选统计特征的数量。

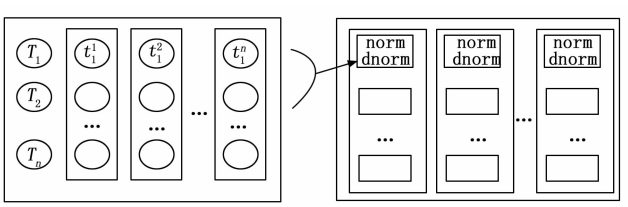

图 2　提取波动特征数据集示意图

3）聚类分析。对于数据集 S，使用 mean-shift 均值漂移的方法进行聚类，针对数据集，随机选择 b 个样本计算其两两之间的距离，并用距离的 c 分位数作为聚类方法所选用的半径（c 分位数即数据中小于等于该数的比例为 c。），将

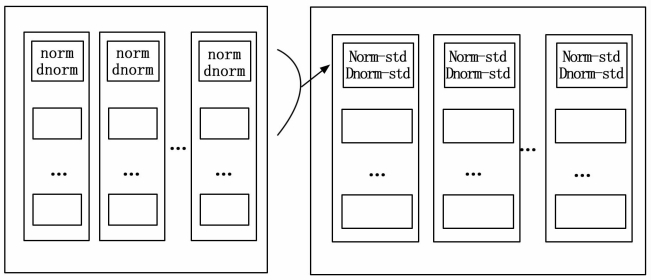

图 3 提取带统计特征的波动特征数据集示意图

高维数据集 S 聚类。b 通常选取 100，c 通常选取 3。

4）异常检测。这里我们选用 iforest 作为算法内层的异常检测手段，其流程分为两个步骤进行：

首先是训练树形模型：从全量数据中抽取部分样本，然后随机选择一个特征作为起始节点，在该特征的最大值和最小值之间随机选择一个值，将样本数据中小于该取值的数据划到左分支，大于等于该取值的划到右分支。接下来在划分好的两个分支数据中不断迭代上述全部步骤，直到满足数据不可再分（只包含一条数据，或者全部数据相同）和二叉树达到限定的最大深度这两个条件时退出，即完成 iTree 构建。

其次是进行模型预测：通过估算它在每棵 iTree 中的路径长度来计算得到所选数据 x 的异常得分。先随机选取一棵 iTree，从根节点开始按不同特征的取值从上往下，直到到达某叶子节点。假设 iTree 的训练样本中同样落在 x 所在叶子节点的样本数为 $T.size$，则数据 x 在这棵 iTree 上的路径长度 $h(x)$，可以用式（9）对其进行计算：

$$h(x) = e + C(T.size) \tag{9}$$

式（9）中，e 表示数据 x 从 iTree 的根节点到叶节点过程中经过的边的数目，$C(T.size)$ 表示在一棵用 $T.size$ 条样本数据构建的二叉树的平均路径长度。通常情况下，我们可以用公式（10）对 $C(n)$ 进行计算：

$$C(n) = 2H(n-1) - \frac{2(n-1)}{n} \tag{10}$$

数据 x 最终的异常分值 Score（x）综合了多棵 iTree 的结果：

$$\text{Score}(x) = 2^{-E(h(x))/C(\varphi)} \tag{11}$$

其中：$E(h(x))$ 表示数据 x 在多棵 iTree 路径长度的均值，$C(\varphi)$ 表示用 φ 条数据构建完成的二叉树的平均路径长度。

从异常分值的公式看，如果数据 x 在多棵 iTree 中的平均路径长度越短，得分越接近 1，表明数据 x 越异常；如果数据 x 在多棵 iTree 中的平均路径长度越长，得分越接近 0，表示数据 x 越正常。

得到上一步骤的聚类结果后，对 S 中的每一类分别使用 Iforest 算法进行异常检测，异常比例 θ 进行统一设置。

5）降维可视化。选取降维算法时要注意到数据线性变化和数据损耗的情况。PCA（principal component analysis），即主成分分析方法，是目前最为广泛应用的数据降维技术。因为 PCA 是一个基于目标数据特征性的最佳正交变换，称它为最佳正交变换主要因为它具备以下较好的特点：转换后与新的能量正交或不有关；转换矢量更趋平衡、能量更趋集中等。PCA 系统由于简单而有效，广泛应用于数据处理中特征选取、数据压缩等各个方面。基于此，本算法中我们选取 PCA 方法将高维数据集 S 降维到 2 维，并进行聚类结果和异常检测结果的可视化。

通过 1）～5），我们便在本算法内部实现了 Iforest 算法和 PCA 算法同滑动窗口的结合，实现了从序列异常到带序列特征的点异常的问题转换。完成了基于滑动窗口的序列异常检测算法的算法设计部分。

3 实验结果及其分析

3.1 数据分析处理

3.1.1 数据特征

通过对实际飞行数据的模拟仿真获取的数据共 2 376 662 条，飞行时间涵盖六个月；对特征的类型进行归类可知主要涵盖直升机架次时间、操控信息和位置信息这三类，特征属性可见表 3。

表 3 直升机数据集特征表

特征类型	特征属性
飞行架次、时间	平台号、时间
操控信息	航速、航向、俯仰、横滚
位置信息	高度、经度、纬度

由于直升机自身的速度限制，不能在很短的时间使得经纬度发生较大的改变，因此我们剔除掉经纬度属性分析，转而使用速度和高度分析空间位置的变化情况。

直升机的姿态角度说明如图 4 所示：航向角为将机体水平方向映射向地面，并与预定目标（一般正北）所形成的夹角，右偏航方向则为正。横滚角表示机翼横轴线与地平面角度，以右倾为正；横滚表示机翼横轴线与地平面夹角，右倾斜为正；俯仰角表示机身纵轴与地平面的角度，以抬头方向为正。

图 4 飞机方向角示意图

3.1.2 数据预处理

通过观察数据文件发现数据中包含一些特殊的特征属性，其经纬度为 0 的异常噪点数据，与实际经纬度显然不符，将这些数据作为噪点数据消除。

经过对数据的观察，经纬度范围相对固定，经度在［110，120］，纬度在［30，40］，因此消除经度在［110，

120］之外、纬度在［30，40］之外的噪点。

3.1.3　异常检测特征选取

为了直观地分析不同特征对于异常检测的影响，我们先对已选择的特征进行可视化分析，由于四维及以上已经超出人脑的空间感受，因此可视化时选择最多三维。

图 5 展示了某一数据集文件中飞行状态中的方向角特征的变化情况，使用 Iforest 异常检测算法设置异常比例为 0.01 进行异常检测，通过 plt 绘制灰度图，正常点为深色，异常点为浅色。从图中可发现直升飞机方向角的数据聚为 4 簇，猜测飞行中存在四种模式，考虑受到机组以及飞行状态的影响存在不同的差别，所以分别获取每个簇对其进行进一步的分析。

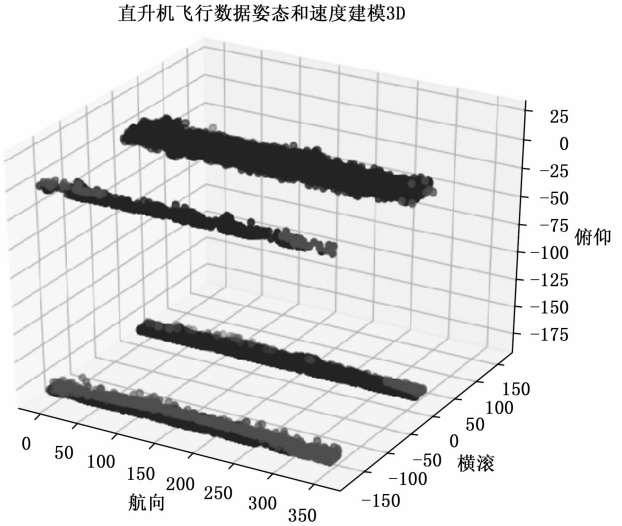

图 5　数据文件中方向角特征异常检测结果

为了研究航线的飞行规律，我们也尝试按照经纬度绘制不同平台飞机的航线，图 6 显示的某一平台上一天采集并记录到的的航线实例。我们使用 pandas 加载飞行数据记录，过滤掉航速为 0 的静止状态记录，提取出飞行轨迹的经度和纬度列表，根据经纬度数据取点连线，绘制了地图和飞行轨迹，飞行轨迹结果见图 6。

图 6　某一平台记录的飞行轨迹图

可以发现，飞机一次飞行的经、纬度改变很小，其它平台也有类似特点，所以经、纬度不作为检测特征。对所有数据进行可视化分析表明飞行状态不随航向发生明显变化，且呈现较为明显的四个簇，每个簇受到飞行平台和机组的不同显现出细微的变化。这与实际也十分契合，因为不同的操作方式不会随着直升机航向的变化而发生变化，只与周围环境以及自身飞行状态有关。飞机姿态由航向角，俯仰角和横滚角确定，航向角仅对航向有影响，因此航向角不作为姿态异常的检测特征。最终选择高度、俯仰角、横滚角和航速特征作为异常检测特征。

3.2　不同航线上的实验验证

3.2.1　算法有效性验证

为了验证算法效率的提升，我们将本算法和上文中提及的几种常用的点异常检测算法应用于网络上几种不同类别的带有异常标签的公开数据集上，进行对比试验。

我们选取了三个不同行业带有不同异常类别的公开数据集，分别为 annthyoid、breastw 和 wbc，针对这三个数据集分别使用 Iforest 算法、PCA 算法和本文提出的滑动窗口算法进行异常检测，并使用 auc 评分进行异常检测效率的对比分析，进行三次试验后，分别对得分取均值作为结果填入表中，结果见表 4。

表 4　多种算法对比 auc 评分

数据集 / 算法	annthyoid	breastw	wbc	平均 auc
Iforest	0.825	0.988	0.940	0.918
PCA	0.673	0.959	0.935	0.856
滑动窗口	0.847	0.975	0.942	0.921

其中横坐标为所用的数据集，纵坐标为所使用的异常检测方法，前两种为滑动窗口中使用到的普适性算法，第三种为本文提出的基于滑动窗口的序列异常检测算法，从结果分析上来看，相较于 Iforest 算法和 PCA 算法，基于滑动窗口的序列异常检测算法有一定的提升，虽然基于滑动窗口的序列异常检测算法仅在两个数据集上跑到了最大值，但是在其他数据集上和其他算法的 auc 评分差距都很小，从而最终相对算法的平均 auc 评分最高。

由此可见，使用基于滑动窗口的序列异常检测算法在算法层面上实现了两个优化，首先是相较于普通的异常检测算法的检测效率有一定的提升；同时将序列异常检测问题转化为点异常检测问题，解决了以往算法不能很好地获取数据间序列特征的问题。

3.2.2　针对聚类结果的预实验分析

为了验证本算法针对无标签数据集的检测效果，我们将其应用于直升机飞行数据集上进行预实验：选取三条航线数据，以 A1-2021-X1-Y1 为例，其表示为 2021 年 X1 月 Y1 日记录在 A1 平台上的全部航线数据。通过对比不同航线聚类结果和异常检测结果来进行验证分析。

所有实验所选择的飞行特征都为高度、俯仰、横滚和航速，且对所有特征都进行归一化处理；窗口均选择参数为 $W_1=5$，步长为 2，$W_2=2$，步长为 1；异常检测算法均使用 Iforest，异常比例设置为 0.02；使用 PCA 将特征数据集 S 降维至 2 维，进行可视化分析。

1）实验一。

数据选择：A1-2021-X1-Y1。

聚类半径选择距离中的 2 分位数，半径为 0.855 1，聚为 4 类。

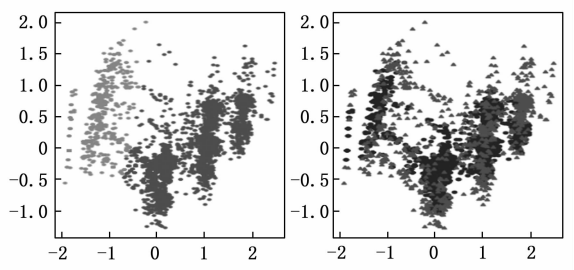

图 7 A1-2021-X1-Y1 聚类结果、异常检测图

2）实验二。

数据选择：A2-2021-X2-Y2。

聚类半径选择距离中的 6 分位数，半径为 1.711，聚为 4 类。

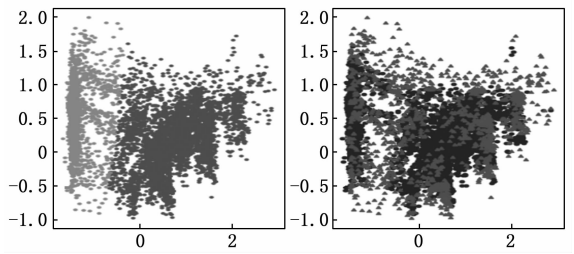

图 8 A2-2021-X2-Y2 聚类结果、异常检测图

3）实验三。

数据选择：A3-2021-X3-Y3。

聚类半径选择距离中的 3 分位数，半径为 0.601 7，聚为 8 类。

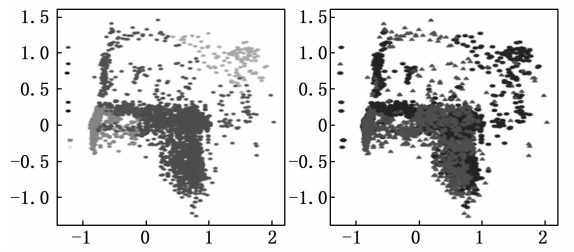

图 9 A3-2021-X3-Y3 聚类结果、异常检测图

上述实验中图 7、图 8 和图 9 中三个左图均为将带有统计信息特征数据集 S 使用上述参数进行聚类后使用 PCA 降维至二维的展示，图中每一种颜色深浅代表高维中聚类的一类。图 7、图 8 和图 9 中三个右图则为将带有统计信息特征数据集 S 中每一类使用 Iforest 进行异常检测并降维可视化至二维平面的示意图，图中圆点为正常数据点，三角为异常数据点。

通过对比三组实验可以发现，通过在高维进行聚类并异常检测后，其降到二维后同一类的点基本在一片区域内，但由于原本维度较大，使用 PCA 降至二维后所选取的特征在二维中不一定是明显的分类，因为损失了部分信息，导致在二维中看起来接近的部分其实是不同的类别。降至低维时其同时筛选出的异常数据均在二维平面所展示类的边缘部分，符合对异常点的定义（即离群点）。同时反映出在这些时间段内，数据的某些或某几个特征存在较大的波动变化。因此从算法上和直观上，都可以展示出该异常检测算法有一定的正确性和可解释性。

最终筛选出存在异常波动的时间窗口。在短时间内，可以认为飞行习惯不会产生较大变化。此时若新增序列加入，可以通过该飞机的历史航线飞行记录对其窗口的统计特征进行聚类，计算出新序列的统计特征，并计算其与各类中心的距离。若距离各个中心都较远，超过某一阈值，则表示该段序列相较原飞行数据可能存在异常，反之则代表其大概率为正常。但若新增序列距离上次聚类时间较长，飞行员的飞行习惯以及直升机的属性等可能发生改变，导致数据聚类的中心发生改变。因此为了保证检测效率，需要使用最近一段的飞行数据重新进行聚类，更新聚类中心。

3.2.3 针对特征变量的实验分析

通过预实验分析，我们验证了基于滑动窗口的序列异常检测算法可以应用于无标签的直升机飞行数据上，接下来将通过飞行特征随时序变化的趋势来验证该异常检测算法的有效性和可解释性。

我们针对试验所选择的飞行特征为高度、俯仰、横滚和航速，将会以对比实验的形式分别，分析以不同飞行特征作为序列异常检测特征时，特征的变化规律、检测为异常时飞行特征是否发生突变等情况，验证算法有效性和可解释性。

通过对实验数据的筛选处理，我们最终选择了 A、B 两条航线通过对比试验的方式对四个飞行特征进行验证。

A 航线：

所选航线为 A4-2021-X4-Y4，其数据条数为 4 352 条。如图 10、11 所示，所选聚类分位数为 3 分位数，聚类半径为 1.235 9，将高维数据聚为 2 类。将其降维后，可以看到异常点几乎分布在每一类的周围。将其中的高度－时间图和俯仰－时间图绘制出来，将异常窗口起始用点标明，并将异常点在灰度图中显示。可以看到，在高度－时间图中，异常点基本处于高度突变且频率较密的地方，同时俯仰角也都在突变区域。由此可见在异常点处其飞行情况确实存在较大波动。

B 航线：

所选航线为 A5-2021-X5-Y5，其数据条数为 6 408 条。如图 12、13 所示，所选聚类分位数为 2 分位数，聚类半径

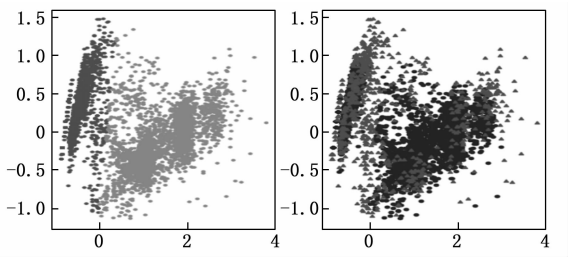

图 10　A4-2021-X4-Y4 聚类结果、异常检测结果图

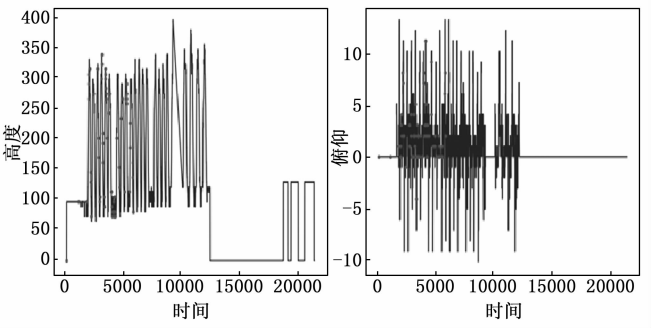

图 11　A4-2021-X4-Y4 高度－时间图、俯仰－时间图

为 1.211 0，将高维数据聚为 2 类。将其降维后，可以看到异常点几乎分布在每一类的周围。将其中的横滚－时间图和航速－时间图绘制出来，将异常窗口起始用点标明，并将异常点在灰度图中显示，可以看到，在航速－时间图中，异常点基本处于高度突变且频率较密的地方，对应到横滚－时间图中可见异常点基本处于高度突变状态中，综合二者可分析得到异常点处其飞行情况确实存在较大波动。

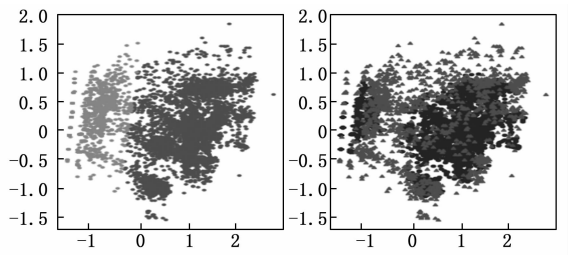

图 12　A5-2021-X5-Y5 聚类结果、异常检测结果

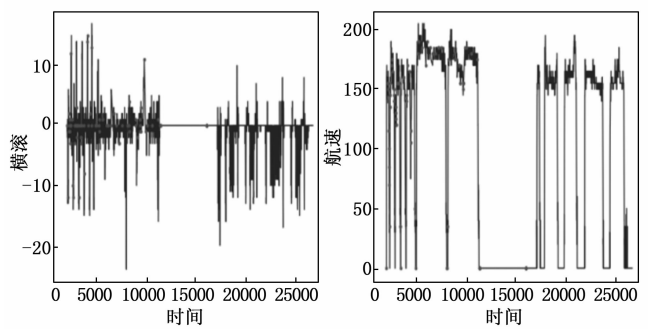

图 13　A5-2021-X5-Y5 横滚－时间图、航速－时间图

由此，可以证明基于滑动窗口的序列异常检测算法不仅在算法层面上对于检测效率有一定的提升，在面向数据层面上也能适应数据集中多维度的特征变量，滑动窗口对于序列特征的数据波动情况和数据统计信息有一定的有效性和可解释性。

4　结束语

为了解决无标签的序列异常检测问题中常常出现的对于数据序列特征获取不充分、无法有效利用；且采用深度学习的方法往往可解释性较差的问题，使用基于滑动窗口的序列异常检测算法来完成针对直升机飞行数据的异常检测：使用多个带标签的公开数据集验证了算法检测效率的提升；并使用无标签数据集验证了算法针对序列异常检测问题的有效性和可解释性。可以为直升机飞行数据等无标签数据的序列异常检测和提供一定的帮助与提升。在后续研究中将针对算法在工业生产中面向实时数据流和数据漂移等情况进一步分析研究。

参考文献：

[1] 侯春萍，赵春月，王致芃，等. 基于有效异常样本构造的视频异常检测算法 [J]. 吉林大学学报（工学版），2021，51（5）：1823－1829.

[2] 范　敏. 外骨骼机器人云脑架构及其学习算法研究 [D]. 成都：电子科技大学，2018.

[3] 苏　昕. 基于数据挖掘的自适应入侵检测系统设计与仿真 [D]. 扬州：扬州大学，2018.

[4] 陈　婧，徐佳琦，李心玥，等. 无监督机器学习异常检测技术在智能监控领域的应用展望 [J]. 中国金融电脑，2021（2）：81－86.

[5] 黄　闯. 基于时序数据挖掘的异常检测系统研究与实现 [D]. 杭州：浙江大学，2021.

[6] 朱海麒，姜　峰. 人工智能时代面向运维数据的异常检测技术研究与分析 [J]. 信息网络安全，2019（11）：24－35.

[7] 邓人博. 基于监视数据的终端区航空器异常行为识别研究 [D]. 天津：中国民航大学，2018.

[8] 黄训华，张凤斌，樊好义，等. 基于多模态对抗学习的无监督时间序列异常检测 [J]. 计算机研究与发展，2021，58（8）：1655－1667.

[9] 刘　静. 多维时间序列异常点检测方法及其应用研究 [D]. 天津：中国民航大学，2020.

[10] 但家梭，马吉林. 基于 PCA-BP 神经网络的船舶动力设备运行状态评价模型 [J]. 船舶工程，2021，43（S1）：357－364.

[11] 张齐家. 基于 PCA 神经网络的电力系统短期负荷预测 [D]. 兰州：兰州交通大学，2017.

[12] 王瑞涵，陈　辉，管　聪. 基于机器学习的船舶机舱设备状态监测方法 [J]. 中国舰船研究，2021，16（1）：158－167.

[13] 王臻睿，赵坤宇，蔡　川，等. 基于 DBSCAN 和 iForest 算法的船舶异常行为分析 [J]. 舰船电子工程，2021，41（4）：89－94.

[14] 田　野. 基于用户用电量的异常检测方法研究 [D]. 重庆：重庆大学，2018.

（下转第 54 页）