一种基于改进后的孤立森林机器学习方法在飞行器控制系统试验电源数据包络分析中的应用

聂　鹏

（北京航天自动控制研究所，北京 100854）

摘　要：为了弥补当前试验数据包络分析方法存在的缺陷，提出并实现了一套新的数据包络分析方法。针对控制系统设计方式相同的不同飞行器多条数据一致性判断问题，采用改良后的孤立森林方法进行数据包络分析，快速找出异常电源电压数据。这种方法还可以推广到类似其他采样数据的数据包络分析场景。在此基础上开发了数据包络分析软件，并进行了多次验证试验。结果表明，提出的数据包络分析方法能有效判断出飞行器控制系统电源所涉及链路中的隐性或显性问题。

关键词：数据包络分析；机器学习；孤立森林

中图分类号：V448.11　　**文献标志码：**A　　**文章编号：**2096-4080（2024）03-0062-06

An Application of an Improved Isolated Forest Machine Learning Method to Envelopment Analysis of Aircraft Control System Power Test Data

NIE Peng

（Beijing Aerospace Automatic Control Institute, Beijing 100854, China）

Abstract: In order to make up for the defects of current test data envelopment analysis methods, a new data envelopment analysis method is proposed and implemented. In order to judge the consistency of multiple data of different aircraft in the same control system, the after improvement isolated forest method is used for data envelopment analysis. Aiming at the consistency judgment of multiple data of different aircraft with the same control system design, the improved isolated forest method is used for data envelopment analysis to find out unqualified power supply voltage data. This method can also be extended to sample data such as otherdata enveloping analysis scenarios. On this basis, a data envelopment analysis software was developed, and several verification tests were carried out. The results show that the data envelopment analysis method proposed in this paper can effectively judge the link problems involved in the aircraft.

Key words: Data envelopment analysis; Machine learning; Isolated forest

0　引言

控制系统[1]是整个航天飞行器的核心，对其产生的试验数据进行研究，可以找出控制系统潜在的问题。由于地面试验有限，能得到产品的试验数据不多，为小样本数据；另外缺乏有效的试验数据包络分析方法，该方法在飞行器测试中是很重要的一环，它决定了试验是否有效，试验的

收稿日期：2024-02-01；**修订日期：**2024-04-03

基金项目：北京航天自动控制研究所质量改进基金

作者简介：聂鹏（1980—），男，硕士，工程师，主要研究方向为控制系统综合及其测试性

正确结论是否能够给出，以及测试中是否出现问题等多个方面[2]。试验数据包络的分析目前主要按区间理论值进行判断。例如控制系统设计方式相同的不同飞行器多条试验数据中得出一个数据簇，统计出最大最小值（作为上下包络值），每一条试验数据都和这个最大值最小值进行比较，得出是否包络的结论。这种方法主要有 3 个缺陷：1）问题数据有可能在最大值和最小值区间内，分析结果易掩盖问题；2）不能有效区分近边缘问题数据；3）对于多条试验数据的一致性，判断不出某套问题设备与其他正常设备数据的不一致性。

本文对同一控制系统批次飞行器的多条电源电压数据进行数据包络分析，可以找出其中的奇异点。通过对奇异点所涉及链路各设备的研究，可以得出数据一致性的好坏，进而发现设备潜在故障，为提前解决设备和飞行器问题奠定基础。

1　样本及算法

1.1　电源电压采样形式

对一二次电源电压进行脉冲采样，形成离散的电源电压数据。飞行器点火前电源电压不稳定，会受带载情况而变化，点火后电压比较稳定，数据就能在一个共同稳定的阶段进行比较，故选择点火后至断电前的稳定电压数据样本进行数据包络分析。电源电压基本曲线图如图 1 所示。

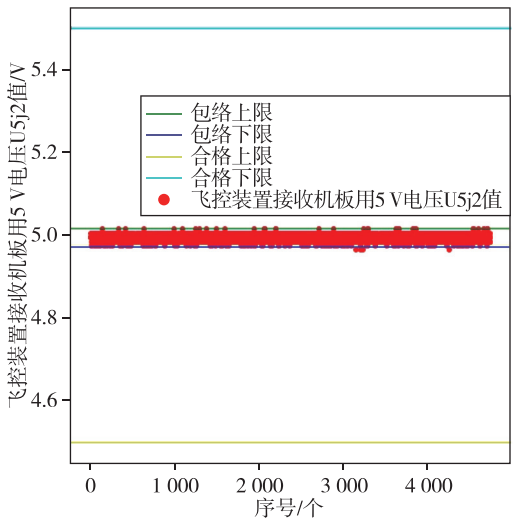

图 1　5 V 电压单条数据的切比雪夫法则 3 个标准差数据图

Fig. 1　The 3-standard deviation range data plots based on Chebyshev's rule for 5 V voltage

1.2　机器学习中的孤立森林方法

通过机器模拟人类的学习行为，并不断获取新的知识信息来刷新已有的知识结构，从而达到不断改善性能的目的。机器学习采用计算机系统，从大量相关数据中自动学习知识信息，不断优化完善算法和统计模型，通过数据集上的误差不断迭代来训练模型，得到对数据集拟合优化的模型，并将训练好的模型用于解决问题。基于训练集中样本输出是否被标记，通常可以将机器学习分为监督学习、半监督学习和无监督学习[3]。

监督学习，也称为监督机器学习，使用标记数据集来训练算法，以便对数据进行分类或准确预测结果。

半监督学习，是监督学习和无监督学习相结合的一种学习方法。半监督学习使用大量的未标记数据和标记数据来进行模式识别工作。

无监督学习，根据类别未知（未标注）的训练样本，解决模式识别中的各种问题。其采用的聚类方法包括 K-Means 方法、Mean-shift 方法、DBSCAN 方法和高斯混合方法等，关联规则方法有 Apriori 算法等。

IForest 算法[4]适用于连续数据的无监督异常检测。与其他异常检测算法相比，IForest 算法通过对原数据集采样建模，减小了数据量太大导致的 masking 和 swamping 效应的影响，算法具备线性时间复杂度和高精准度，是符合当前大数据处理要求的 state-of-the-art 算法[5]。

由于本文面临的数据包络分析问题没有标记数据，无法进行有效训练，故选择非监督学习中的孤立森林方法[6]来寻找孤立点。

2　数据包络分析

2.1　孤立森林方法原理

2.1.1　构建孤立森林

1）从数据集 D 中随机选择 m 个样本点作为生成单棵孤立二叉树的样本集 Sd。

2）从样本集 Sd 中随机选择一个特征 f 和一个切割值 p。若结点 N 包含的所有样本在特征 f 下的最大值和最小值分别为 $f_{\max}$ 和 $f_{\min}$，则有 $p[f_{\min}, f_{\max}]$。

3）若样本关于特征 f 的值小于切割值 p，则将该样本划分到结点 N 的左孩子。否则，划分到右孩子。

4）重复 2）、3）两步，分别对结点 N 的左右孩子结点进行分割，生成孤立二叉树。当孩子结点中有多条相同的数据或只有一条数据或孤立二叉树已到达设置的最大高度时，停止生成。

2.1.2 计算测试样本点的异常分值

根据测试样本点 d 在各孤立二叉树中的路径长度 $h(d)$，利用公式计算 d 的异常值，从而评价其异常情况。路径长度越小，则该样本点的异常分值越高。

2.2 孤立森林方法进行数据包络分析

从图 2 中可以看出，红色为合格包络数据，蓝色为合格不包络数据，从图中可见有部分合格不包络的数据被认为合格包络了，在图像右下方夹在蓝色数据中间的部分。

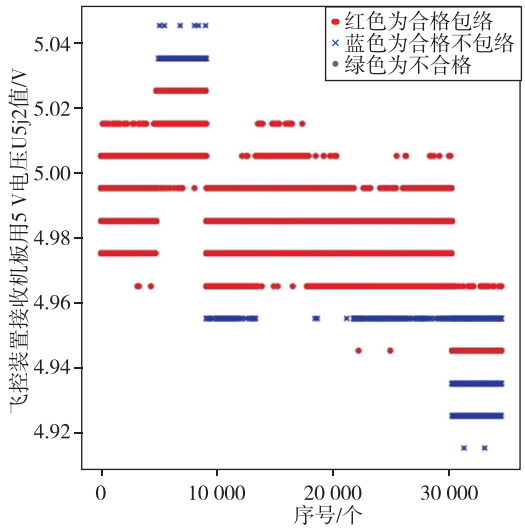

图 2 5 V 电压 8 条数据的孤立森林算法数据包络分析

Fig. 2 The data envelope analysis of isolated forest based on 8 data of 5 V voltage

2.3 原孤立森林方法中以 96%的概率统计确定孤立点进行数据包络分析

试验以 89%的概率统计出所有合格和不合格的点，测试结果不稳定。以 96%的概率统计出所有合格和不合格的点，然后确立每条数据中这些点的位置，定位到单条数据。当数据为基本正确时算法效果明显，但孤立点的数量变大时，算法受到干扰，会将明显的孤立点误判成合格点，见图 3。

该算法将部分奇异点异常数据都识别成正常数据（图左数据蓝色点中的红色部分）。

2.4 改进孤立森林方法进行数据包络分析

基于自动寻值直方图孤立森林改进算法，通过自动寻值给程序喂值，在大范围参数中找到最好的程序配对值，用直方图方法统计出所有合格和不合格点，然后在每条数据中确立这些点的位置，定位到单条数据。从试验结果（见图 4）看，改进方法能很好地定位孤立点，明确试验数据是否具有一致性。

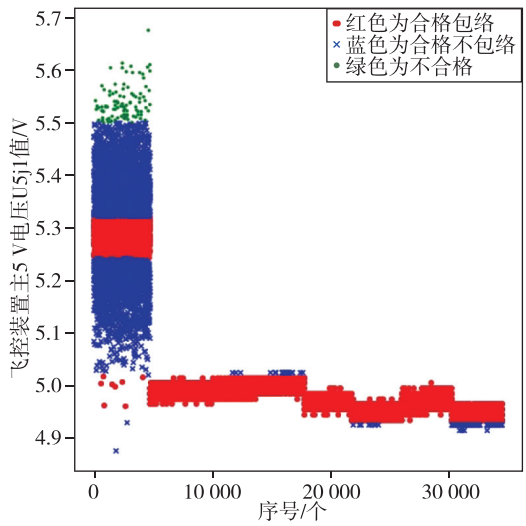

图 3 5 V 电压 8 条数据包络分析图

Fig. 3 5 V voltage 8 data envelop analysis images

如图 4 所示，第一条数据（全体被标为蓝色和绿色的数据）由合格不包络数据和不合格数据组成，故第一条数据不具有一致性。

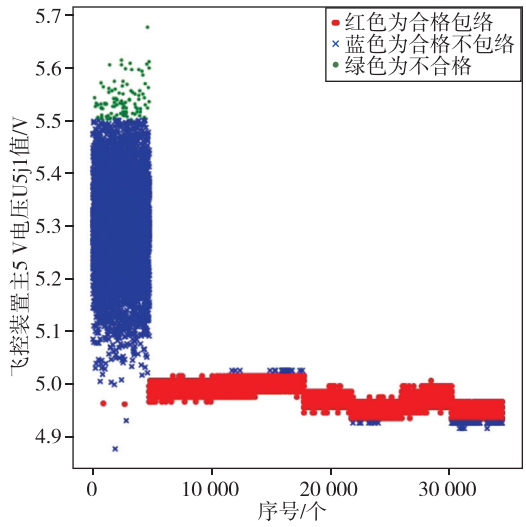

图 4 5 V 电压 8 条试验数据的改进孤立森林算法数据包络分析图

Fig. 4 Improved isolated forest algorithm performs data envelopment analysis images based on 8 test data of 5 V voltage

对于 8 条数据中一致性强的数据，可以分析出

最差电源链路飞行器。

1）对同一控制系统批次飞行器的数据进行纵向分析，结果见图 5。将孤立点定位到每一条试验数据，结果见表 1。

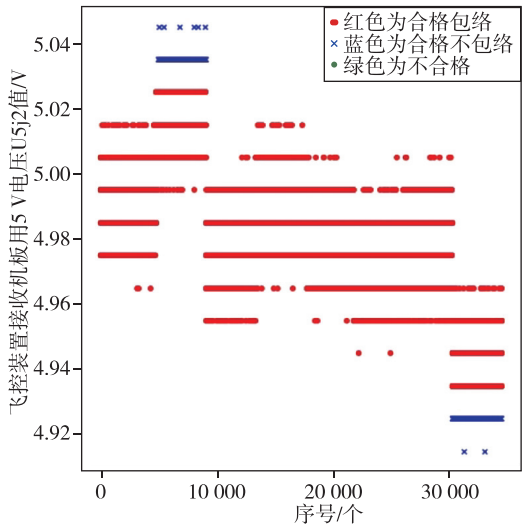

图 5　改进的孤立森林方法数据包络分析图

Fig. 5　Improved data envelopment analysis diagrams for isolated forest method

表 1　孤立点定位到每一条试验数据的试验结果

Tab. 1　Isolated points of each test data to get the test results

数据	总数/条	不合格数/条	不合格占比/%
第 1 条	4 757	0	0
第 2 条	4 287	218	5.085
第 3 条	4 291	42	0.979
第 4 条	4 431	0	0
第 5 条	3 993	3	0.075
第 6 条	4 261	106	2.488
第 7 条	4 221	42	0.995
第 8 条	4 281	2 019	47

从数据可以看出，第 1 条和第 4 条数据最好，第 8 条数据效果最差。

2）电压采样数据横向分析结果见表 2～表 4。从表中数据可以分析出，母线电压稳定性相对稍差。第 1，2，5 条飞行器控制系统电压数据最稳定，第 3，4，6，7 条控制系统电压数据相对稳定，而第 8 条控制系统电压相对稳定性较差。

表 2　各电压孤立点占比值表

Tab. 2　Isolated points with the total number of points

序号	5 V 电压 1	5 V 电压 2	电池电压	某电压	母线电压	变换电压	稳压电压
1	0	0	0.05	0	0.18	0	0
2	0.05	0	0.15	0.1	0	0	0
3	0.1	0.01	0.21	0.06	0	0	0
4	0	0.2	0.03	0	0.24	0.1	0.09
5	0	0	0.05	0	0	0.03	0.08
6	0.02	0.07	0.02	0.41	0	0	0.17
7	0	0	0.03	0.02	0.25	0	0.13
8	0.47	0.2	0.15	0	0	0.56	0

表 3　各电压孤立点占比值表（按阈值 0.2 及以下都为 0）

Tab. 3　Isolated points with the total number of points（0 if the threshold value is 0.2 or less than 0.2）

序号	5 V 电压 1	5 V 电压 2	电池电压	某电压	母线电压	变换电压	稳压电压
1	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0
3	0	0	0.21	0	0	0	0
4	0	0	0	0	0.24	0	0
5	0	0	0	0	0	0	0
6	0	0	0	0.41	0	0	0
7	0	0	0	0	0.25	0	0
8	0.47	0	0	0	0	0.56	0

表 4 各电压孤立点占比值表（剔除 0 值）

Tab. 4 Isolated points with the total number of points (eliminated by 0 value)

序号	5 V 电压 1	电池电压	某电压	母线电压	变换电压
1	0	0.21	0	0	0
2	0	0	0	0.24	0
3	0	0	0.41	0	0
4	0	0	0	0.25	0
5	0.47	0	0	0	0.56

2.5 孤立森林方法验证

2.5.1 切比雪夫法则

利用切比雪夫法则对同一控制系统批次飞行器的不同发次 7 种电源电压 8 条总数据进行一致性的数据包络分析，如图 6 所示。

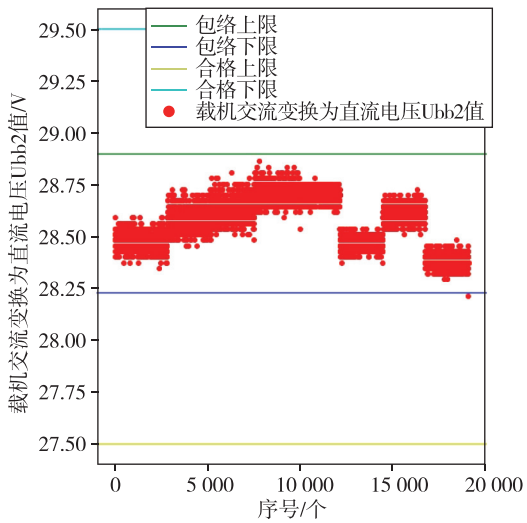

图 6 利用切比雪夫法则进行一致性的数据包络分析表

Fig. 6 Chebyshev's rule for consistent data envelope analysis

从图像来看，上下包络线无法把 8 条数据的一致性体现出来，数据间变化小则能正确反映，一旦数据变化大，包络具有随形化趋势，无法正确反映数据的一致性。

2.5.2 机器学习 K-Means＋＋方法

K-Means＋＋方法需要人工绘制肘部图识别其中数据的拐点，来作为最好的聚类值，从而对数据进行分类。运用轮廓算法自动算出最好的聚类值点，对数据进行分类。当发生数据有少数条目和大多数条目不一致时，能够将完好的条目聚为一类，然后少数条目独立聚类，从而找到数据不一致的条目。运用轮廓算法自动算出最好的聚类值点，在聚类为 8 时打分值最高 0.96。图 7 为采用 K-Means＋＋方法的聚类图（聚类为 8）。

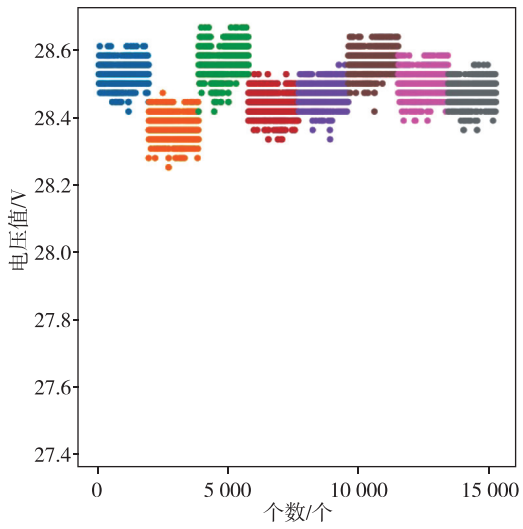

图 7 K-Means＋＋方法对数据进行聚类

Fig. 7 K-Means＋＋ method to cluster the data

2.6 孤立森林方法准确性验证

利用随机数据尽可能逼近实际数据进行准确性测试。表 5 为改进孤立森林测试准确率。从表中可以看出，数据信噪比在 2%～6%之间准确率最高，数据信噪比在 10%～14%之间准确率也容易达到最大值。

表 5 改进孤立森林测试准确率

Tab. 5 Improved isolated forest test accuracy table

序号	[异常电压均值/V，标准差/V，随机数/个]	准确率/%	信噪比/%
1	[20，0.2，100]	0.9	0.5
2	[29，0.2，100]	0.9	0.5
3	[35，0.2，100]	0.7	0.5
4	[40，0.2，100]	0.6	0.5
5	[20，0.2，400]	0.8	2
6	[29，0.2，400]	0.8	2
7	[35，0.2，400]	0.9	2
8	[40，0.2，400]	0.8	2
9	[20，0.2，1 000]	0.85	5
10	[29，0.2，1 000]	0.95	5
11	[35，0.2，1 000]	0.85	5
12	[40，0.2，1 000]	0.75	5
13	[20，0.2，1 600]	0.67	8
14	[29，0.2，1 600]	0.67	8
15	[35，0.2，1 600]	0.97	8
16	[40，0.2，1 600]	0.67	8

续表

序号	[异常电压均值/V，标准差/V，随机数/个]	准确率/%	信噪比/%
17	[20，0.2，2 000]	0.84	10
18	[29.5，0.2，2 000]	0.89	10
19	[30，0.2，2 000]	0.79	10
20	[40，0.2，2 000]	0.69	10
21	[20，0.2，2 500]	0.71	12.5
22	[29.5，0.2，2 500]	0.81	12.5
23	[30，0.2，2 500]	0.81	12.5
24	[40，0.2，2 500]	0.81	12.5
25	[20，0.2，2 800]	0.8	14
26	[29.5，0.2，2 800]	0.9	14
27	[30，0.2，2 800]	0.7	14
28	[40，0.2，2 800]	0.9	14

2.7　软件实现

数据包络分析程序软件界面见图 8。编程方法中使用了多线程的方式，大大缩短了单一线程数据处理运行的时间。

图 8　数据包络分析程序软件界面

Fig. 8　Data envelope analysis program software interface

试验数据包络分析方法的中间图像和数据会在软件平台上显示，见图 9。

2.8　结论

在传统飞行器数据包络分析中引入改进孤立森林方法，解决了原方法的 3 个问题，并能得出有效的数据问题结论。试验结果表明，对于现有的数据，能有效判断出飞行器某套问题设备与其他正常设备数据的不一致性，对进一步排查设备问题提供了有效可靠的依据。

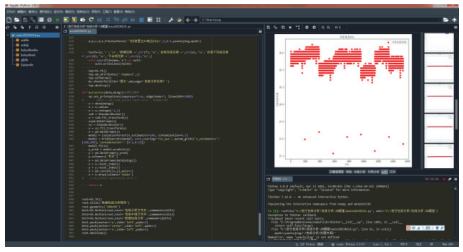

图 9　软件平台上显示的图像和数据文件

Fig. 9　Software platform displays images and data files

3　结束语

在数据包络分析中，基于人工智能算法，制作数据自动分析工具，利用统计学和机器学习等方法，进行数据包络分析研究，科学性发现设备问题，为提前解决飞行器问题奠定理论基础。后续将集中于试验数据的差异性分析，通过试验数据分析进一步找出设备的使用瓶颈，确认设备的健康状态。

参考文献

[1] 徐延万. 控制系统(上)[M]. 北京:中国宇航出版社，2005：2-34.

[2] 曾竹喧. 运用数据包络分析方法的机场群效应研究[D]. 南京：南京航空航天大学，2020.

[3] 白杨. 机器学习的隐私关键问题研究 [D]. 成都：电子科技大学，2022.

[4] Liu F T，Ting K M，Zhou Z H. Isolation-based anomaly detection[J]. ACM Transactions on Knowledge Discovery from Data，2012，6(1)：1-39.

[5] 左凡秀. 基于信息熵的孤立森林算法改进和并行实现[D]. 汕头：汕头大学，2020.

[6] 朱丽琴. 基于孤立森林的入侵检测方法研究 [D]. 哈尔滨：哈尔滨工程大学，2020.

引用格式：聂鹏.一种基于改进后的孤立森林机器学习方法在飞行器控制系统试验电源数据包络分析中的应用[J].宇航总体技术，2024，8(3)：62-67.

Citation：Nie P. An application of an improved isolated forest machine learning method to envelopment analysis of aircraft control system power test data [J]. Astronautical Systems Engineering Technology，2024，8(3)：62-67.