Journal of Intelligent & Fuzzy Systems 34 (2018) 3785–3798
DOI:10.3233/JIFS-169551
IOS Press

Intelligent pattern analysis and anomaly detection of satellite telemetry series with improved time series representation

Jingyue Pang, Datong Liu*, Yu Peng and Xiyuan Peng
Department of Automatic Test and Control, Harbin Institute of Technology, Harbin, P.R. China

Abstract. Telemetry data, sent by the satellite, is the only basis for ground staffs to monitor on-board equipment status. In addition, the pattern discovery and operating state identification of telemetry data are very essential for automatic anomaly detection and problem diagnosis for satellites. Clustering, as an important data mining method for time series, can realize pattern discovery of satellite telemetry data automatically and intelligently, whereas the large amount of raw data and pseudo-period characteristic make the clustering on raw data inefficient and susceptible to noise interference. Thus, based on the prominent shape features and Time-Spatial specialty, a clustering framework is proposed for telemetry data mining with physical-based segmentation and improved time series representation. Moreover, different distance measures are introduced to this framework to realize the time series clustering. The experiments are firstly performed on the public data sets which have high similarity with the real satellite telemetry to quantify the clustering accuracy, then a case study on the real satellite telemetry verifies the effectiveness and applicability of the proposed framework.

Keywords: Satellite telemetry time series, clustering, representation, special-points series

1. Introduction

Telemetry data is the only basis for ground staffs to judge the working performance and health status of an on-orbit satellite [1, 2]. During a satellite working in orbit, a large amount of telemetry data is collected and stored in the ground database, but only a few of them which beyond the threshold can be applied for further analysis. Especially with the development of acquisition and sensor technique, the amount of one satellite telemetry has reached to the level of TB. Consequently, a great unbalance has appeared between storage resources and latent value [3]. Therefore, it is necessary to develop the intelligent and automatic tools to mine the latent value within the satellite telemetry data [4]. It has great significance in providing effective support for state monitoring, fault diagnosis and health management of satellites [5–7].

Among the data mining methods, clustering analysis is an effective and intelligent measure to identify the intrinsic structure, relations, and interconnectedness of the complicated satellite telemetry data without expert experience [8, 9]. In addition, it provides effective information for further subsequent research, including classification [10, 11], prediction [12], association rule discovery, anomaly detection [13, 14], and so on. Clustering methods generally refer to partitioning-based clustering, hierarchical clustering, density-based clustering, grid-based clustering, and model-based methods [15]. Considering the advantages of easy-to-definite and fewer constrictions on the number of cluster [16], agglomerative

*Corresponding author. Datong Liu, Department of Automatic Test and Control, Harbin Institute of Technology, Harbin 150080, P.R. China. E-mail: liudatong@hit.edu.cn.

hierarchical clustering (AHC) is the focus of this work.

However, satellite telemetry data is time series with the typical characteristics of long time range and large amount. Clustering on original series is inefficiency and sensitive to noise. Therefore, it is very necessary to segment the telemetry and create an effective representation of the raw series as the input of clustering to realize high performance clustering.

A fixed window or time range should be considered firstly to realize telemetry series segmentation. While, many telemetry series appear highly periodic, but never exactly repeats itself. Specifically, the telemetry data constitute a pseudo-periodic time series. In this case, the fixed window-based segmentation will cause poor synchronicity which further influences the clustering accuracy. Thus, the orbit argument which reflects the Time-Spatial specialty is employed to segment the raw series to obtain the improved subseries of each orbit period.

On the other hand, five methods are usually implemented for time series representation, i.e., singular value decomposition (SVD) [17], frequency-domain transformation [18], piecewise linear representation (PLR) [19], model-based method [20] and symbolic representation [21]. It is worth mentioning that, due to the requirements of higher compute efficiency and lower characteristic distortion, PLR is more suitable and effective for telemetry data representation.

PLR-based methods can be conducted with piecewise aggregate approximation (PAA) and feature points. PAA is one of the most popular methods for time series dimensionality reduction and feature extraction [22]. This method divides original time series into equal length series and extracts the main information (such as mean, variance and slope) of each subsequence to create a feature series [23, 24]. PAA is applied to a wide range of data types and is simple to calculate. However, due to the regularity of most satellite telemetry data, this method is likely to cause some key information and important patterns missing. PLR-based methods with feature points extract the special points, such as extreme value points and key turning points to represent the raw series [25]. Special-points series contains more meaningful pattern information which is effective for improving the subsequent clustering accuracy. Whereas, this method only pays attention on the turning points between mode changes without considering the smooth modes, consequently, the clustering accuracy is influenced with the input of the distortion series.

Hence, this paper presents an improved time series representation method based on Special Points Series Segmentation (SPSegmentation). SPSegmentation shows two aspects of improvements on extending the definition of local extreme value and optimizing the holding time of local extreme value, so that it can extract the most representative samples. However, two different special-points series generated by SPSegmentation may face the problems of unequal length and time asynchronization. The unequal series limit the available types of measure function. In order to solve this problem, a processing of isometric treatment in pairs is necessary. Based on isometric extension, the previous special-points series can be measured by more distance measures.

Finally, based on Time-Spatial specialty and the improved time series representation, a clustering framework of satellite telemetry is proposed combined with AHC algorithm. Given that there are no actual category labels within the satellite telemetry time series, this paper firstly adopts the open datasets which have high similarity with the satellite telemetry data to quantify the clustering performance. Then clustering for actual telemetry data is realized with the proposed framework, some analysis on the actual telemetry series verify its effectiveness and applicability.

This paper is organized as follows. Section 2 introduces the AHC clustering based on time series representation, especially, SPSegmentation and isometric treatment are described in this part. In Section 3, the proposed framework of telemetry clustering is discussed in detail where much analysis is made on the real telemetry data. Experimental results on the public data sets and actual satellite telemetry data will be presented in Section 4. Finally, Section 5 concludes the work and describes some future work.

2. Time series clustering based on improved time series representation

2.1. *Time series representation based on key points segmentation*

Key points, generally referring to turning point and local extreme points, are selected for representing the raw time series. Where a turning point, shown in Fig. 1, meets the consideration of Equation (1).

$$b = \left| x_i - \frac{x_{i-1} + x_{i+1}}{2} \right| \geq \varepsilon \tag{1}$$

where ε is the setting distance threshold [26].

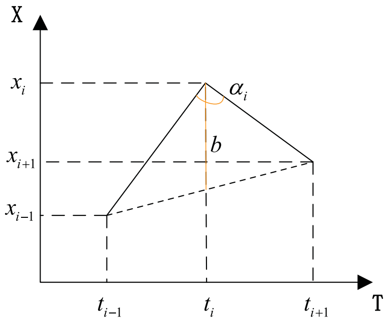

Fig. 1. Minimum sequence constituted by three continuous points.

In addition, a local extreme point meets the constraint condition of Equation (2).

$$\begin{cases} x_i < x_{i+1} \\ x_i < x_{i-1} \end{cases} \text{or} \begin{cases} x_i > x_{i+1} \\ x_i > x_{i-1} \end{cases} \tag{2}$$

For time series $X = \{x(t_1),\ x(t_2), \ldots,\ x(t_n)\}$, time series representation based on Key Points Segmentation (KPSegmentation) extracts the important turning points and local extreme points to represent the raw series. The detailed description is shown as follows.

Firstly, the initial local extremum series, IM, is formed by two endpoints $x(t_1)$, $x(t_n)$ and other points meeting Equation (3).

$$\begin{aligned} &\{x(t_i) < x(t_{i-1}) \cap x(t_i) < x(t_{i+1})\} \\ &\quad \cup \{x(t_i) > x(t_{i-1}) \cap x(t_i) > x(t_{i+1})\} \end{aligned} \tag{3}$$

where $2 \le i \le n - 1$; $IM = \{x(t_{p_j})\}_{j=1}^{m}$, $m \le n$; $p_1 = 1$ and $p_m = n$.

Then IM series is filtered by Equation (4).

$$p_{j+1} - p_{j-1} > C \tag{4}$$

where $2 \le j \le m - 1$, C is the setting holding time. And the points which meet Equation (4) are added into a new extremum set, M. denoted by $M = \{x(t_{q_j})\}_{j=1}^{k}$, where $k \le m$; $q_1 = 1$ and $q_k = n$.

On the other hand, the turning points series, N, is made up by the points in X meeting Equation (5).

$$\left| x(t_i) - \frac{x(t_{i+1}) - x(t_{i-1})}{2} \right| \ge \varepsilon \tag{5}$$

where $1 < i < n$, $N = \{x(t_{v_j})\}_{j=1}^{h}$, $h \le n$.

Finally, sort all of these points from series M and N in the order of time to form the key-points series, and remove duplicate data.

With these key points, the original time series can be fitted linearly while some small portion of noise is reduced. However, in the dense extreme areas, the extremum may be continuously deleted with the holding time of the extremum much larger than C.

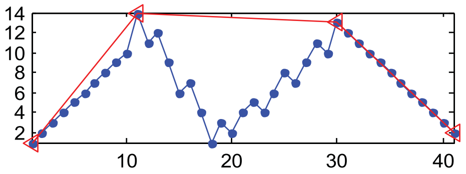

Fig. 2. An example of KPSegmentation ($C = 3$, $\varepsilon = 3$).

An example is shown in Fig. 2 that we set the holding time much smaller than the general setting. As a result, the local extremum between 11 and 30 are all deleted when C is set to 3, in detail, the time interval between adjacent key points at time index 11 and 30 are 19, much larger than 3.

In order to solve this problem, this work introduces an improved KPSegmentation method, named SPSegmentation, to extract special-points series (SPS).

2.2. *Improved key points segmentation for time series representation*

In this work, we improve the definition of local extreme value as well as holding time of the KPSegmentation method. On one hand, SPSegmentation extends the local extremum definition by adding the turning points between smooth modes and other modes. On the other hand, it optimizes the selection strategy of local extreme value so that the time interval of local extremums can be kept closed to C after filtering.

Improvement 1. Extend the definition of turning points to get the improved series IM which is formed by these points meeting Equation (6).

$$\begin{aligned} &\{x(t_i) \le x(t_{i-1}) \cap x(t_i) < x(t_{i+1})\} \\ &\quad \cup \{x(t_i) < x(t_{i-1}) \cap x(t_i) \le x(t_{i+1})\} \\ &\quad \cup \{x(t_i) \ge x(t_{i-1}) \cap x(t_i) > x(t_{i+1})\} \\ &\quad \cup \{x(t_i) > x(t_{i-1}) \cap x(t_i) \ge x(t_{i+1})\} \end{aligned} \tag{6}$$

As shown in Fig. 3, KPSegmentation only selects the turning points of a and b, while SPSegmentation also adds the points of c, d, e, and f.

Improvement 2. Optimize the selection strategy of local extremum. SPSegmentation improves the holding time as follows with the setting holding time as C.

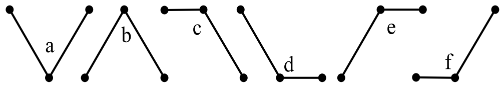

Fig. 3. Basic modes of turning points.

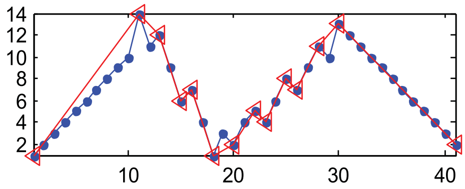

Fig. 4. Result of SPSegmentation ($C = 3$, $\varepsilon = 3$).

```
$IM = \{x(t_{p_j})\}_{j=1}^{m}$
$j = 2,\ m = \text{length}(IM)$
while $j < m-1$ {
  if $(p_{j+1} - p_{j-1} \leq C)$
    Delete $x_{p_j}$ from $IM$
    $m = \text{length(IM)}$
  else
    $j = j + 1$}
```

By using the same testing series with Fig. 2, the result of the improved SPSegmentation method is shown in Fig. 4. It is obvious that this method effectively solves the problem of continuous deletion, and the selected time interval between these local extremes is mostly close to the holding time threshold of C.

In general, for the given time series $X = \{x(t_1),\ x(t_2), \cdots,\ x(t_n)\}$, holding time of local extremum C and the threshold of triangle's midline distance ε should be set in advance, then the special-points series $\text{SPS} = \{x(t_{s1}),\ x(t_{s2}), \cdots,\ x(t_{sv})\}$ can be obtained by SPSegmentation. Undoubtedly, the performance of SPSegmentation is affected by the setting parameters. While feasible setting parameter generally needs the preliminary data analysis and parameter pre-judgement of the dataset. Therefore, this paper provides an adaptive method for obtaining effective parameter ε. The detailed process is described as follows.

- First, calculate the reference turning coefficient for all points in the whole sequence by Equation (7), except the starting point and the ending point.

$$z = \left| x(t_i) - \frac{x(t_{i+1}) - x(t_{i-1})}{2} \right| \tag{7}$$

- Then, calculate the upper quartile (Q_1) and lower quartile (Q_3) of sequence z. The upper quartile, also called 1st quartile, is the point that splits off the highest twenty-five percent of the series from others. Similarly, the lower quartile, also called 3rd quartile, splits off the lowest twenty-five percent of the series from others. Then, calculate the interquartile range IQR by Equation (8).

$$\text{IQR} = Q_1 - Q_3 \tag{8}$$

- Calculate the abnormal cut-off point according to the upper quartile and the interquartile range by Equation (9), the result of which may be set as the parameter ε.

$$\varepsilon = Q_3 + 1.5 \times IQR \tag{9}$$

After this parameter is determined, users can control the holding time of local extremums to improve the compression rate and quality of SPS.

2.3. Isometric treatment in pairs and similarity measure

Corresponding special-points series of each subsequence are derived based on SPSegmentation. Then select any two-different special-points series SPS_i of Equations (10 and 11).

$$SPS_i = [x_i(t_{p1}),\ x_i(t_{p2}), \cdots,\ x_i(t_{pk})] \tag{10}$$

$$SPS_j = [x_j(t_{q1}),\ x_j(t_{q2}), \cdots,\ x_j(t_{qm})] \tag{11}$$

where $m < n,\ k < n,\ t_{p1} = t_{q1} = t_1,\ t_{pk} = t_{qm} = t_n$; and generally, $m \neq k,\ t_{pv} \neq t_{qv}$ which is shown in Fig. 5.

Obviously, the quantity of special points and the corresponding position of time axis may not be the same for different SPSs. Since it is hard to measure similarity distance between two unequal series, isometric treatment is necessary. Hence, before the procedure of distance measure, isometric treatment will be performed on each two different SPSs, so that the time axis of two SPSs are aligned.

In detail, take the union of two subscript sets $\{p_1,\ p_2, \cdots,\ p_k\}$ and $\{q_1,\ q_2, \cdots,\ q_m\}$, and sort it in ascending order to obtain the set $\{v_1,\ v_2, \cdots,\ v_h\}$, where $\max(m,\ k) \leq h \leq m + k - 2$. According to the union set, we finally extract two Pairwise Special Points Series (PSPS), shown in Equations (12 and 13).

$$PSPS_{ij} = [x_i(t_{v1}),\ x_i(t_{v2}), \cdots,\ x_i(t_{vh})] \tag{12}$$

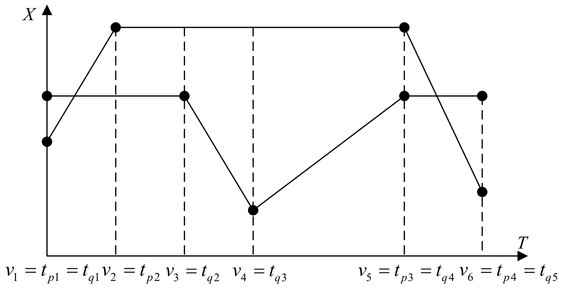

Fig. 5. Isometric treatment in pairs.

$$PSPS_{ji} = [x_j(t_{v1}),\ x_j(t_{v2}), \cdots,\ x_j(t_{vh})] \quad (13)$$

After isometric extension, the similarity between two PSPSs can be calculated by various distance measures, such as Euclidean distance, correlation coefficient distance and Dynamic Time Warping (DTW).

$$\begin{aligned} & PSPS\_dist\,(SPS_i, SPS_j) \\ & \quad = dist(PSPS_{ij}, PSPS_{ji}) \end{aligned} \quad (14)$$

where *dist* represents any distance measure. In this work, we choose Euclidean distance, Spearman correlation distance, DTW distance and Derivative Dynamic Time Warping (DDTW) distance for experimental verification.

2.4. *AHC based on improved Special points series selection*

AHC is a bottom-up clustering method, which agglomerates the closest pair of clusters by similarity measure [27, 28]. The framework of time series clustering based on the SPSegmentation method, isometric treatment and AHC algorithm is shown in Fig. 6.

Step 1. Preprocess the raw time series set $X = [X_1,\ X_2, \cdots,\ X_l]$ with z-score algorithm.

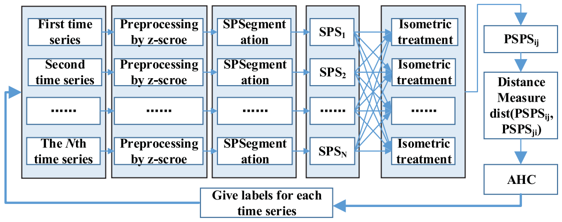

Fig. 6. AHC with SPSegmentation and isometric treatment.

Step 2. Based on SPSegmentation method, extract all special points from each raw time series X_i to form a special-points sequence $SPS_i = \left[x_i\left(t_{p_1}\right),\ x_i\left(t_{p_2}\right), \cdots,\ x_i\left(t_{p_k}\right)\right]$ as the input of the clustering method.

Step 3. Isometric treatment in pairs for every two special-points series, SPS_i and SPS_j to obtain the series of PSPS_{ij}.

Step 4. Measure similarity distance of every two series of PSPS_{ij} and obtain the similarity coefficient matrix.

Step 5. Combined with AHC algorithm to realize time series clustering, and acquire the label of each time series.

3. Proposed framework for telemetry series clustering

3.1. *Telemetry series analysis*

Satellite is generally launched into orbit around one planet to perform certain tasks. For example, the satellite in Sun-synchronous orbit passes over any given point of the planet's surface at the same local solar time [29]. And the surface illumination angle will be nearly the same every time that the satellite is overhead. Whereas a geosynchronous orbit (sometimes abbreviated GEO) is an orbit around Earth of a satellite with an orbital period that matches Earth's rotation on its axis [30], which takes one sidereal day (23 hours, 56 minutes, and 4 seconds). It is obvious that an orbit of a satellite is relatively regular, repeating path. In addition, the working conditions show the periodic change, such as the change of charging and discharging mode in power subsystem. Therefore, there are many analog telemetry series shown the pseudo-periodic phenomenon as shown in Fig. 7.

3.2. *Telemetry series segmentation*

Analog telemetry series with pseudo-periodic characteristic are the focus of this work. In order to realize intelligent pattern discovery, the raw series need to be segmented into some subseries. Obviously, the periods of these series are firstly applied for subseries segmentation. Take one angle series as an example which is shown in Fig. 8a) and 8b) shows the segmentation result based on the fixed period calculated by FFT analysis.

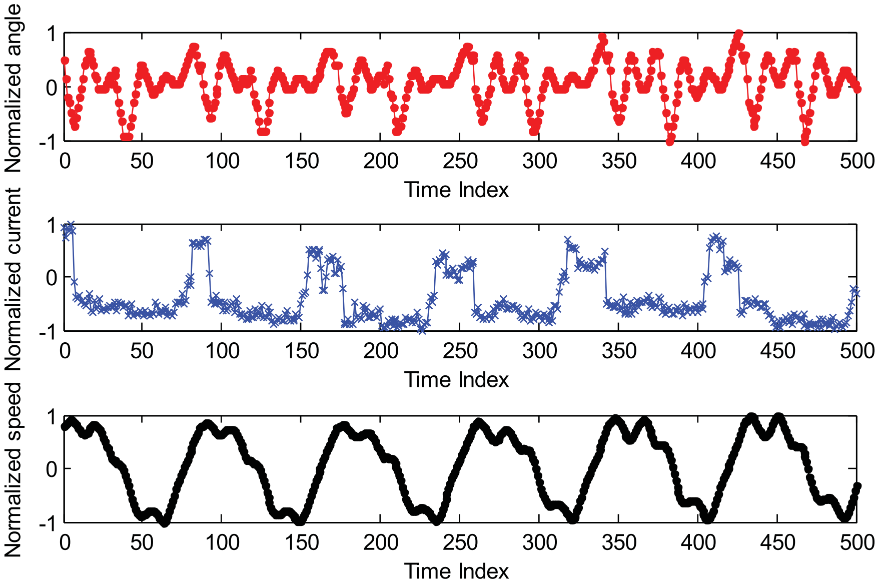

Fig. 7. Examples of satellite telemetry series.

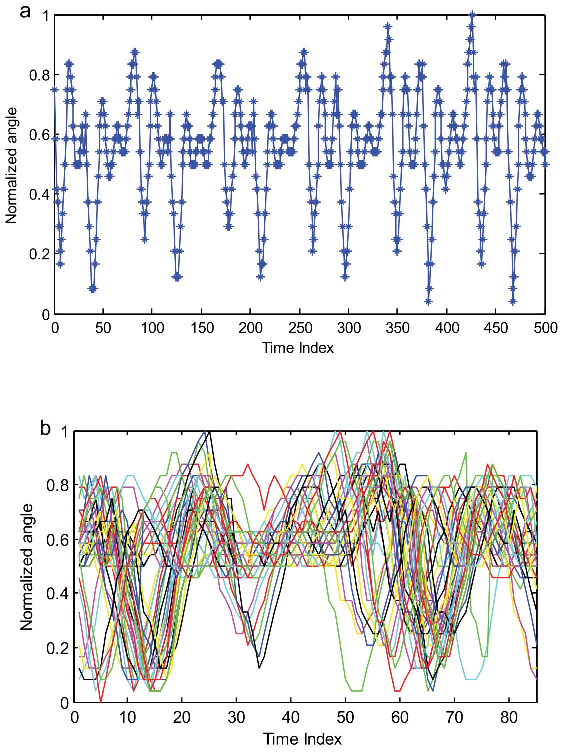

Fig. 8. a) The raw angle telemetry series. b) Angle subseries based on fixed period segmentation.

As shown in Fig. 8b), the subseries segmented by the fixed period of 86 are not consistency, the reason is that the angle series is not a strict periodic series. Hence, the Time-Spatial specialty is considered to segment raw series, in detail, the orbit augment, ranging from 0 to 360 degrees, is selected to segment raw series. And it has high correlation with the analog series shown in Fig. 9a). The subseries segmented by orbit argument are shown in Fig. 9b).

As shown in Figs. 8b) and 9b), these subseries segmented by orbit argument have a higher consistency which provides an important basis for improving the subsequent clustering accuracy.

3.3. AHC algorithm framework for telemetry series

Based on the analysis and discussion on telemetry series and segmentation method described in Sections 3.1 and 3.2, the AHC framework for satellite telemetry series is proposed with SPSegmentation and isometric treatment as shown in Fig. 10.

1) Extract an analog series and the orbit argument series from the history telemetry database with the setting time interval. It is noted that some prepressing operations are performed in this step referring to missing values processing, outliers removing and data standardization, etc.
2) Take the orbit argument changing from 0 to 360 degrees as a period to segment the analog series. Some subseries can be derived in this procedure.

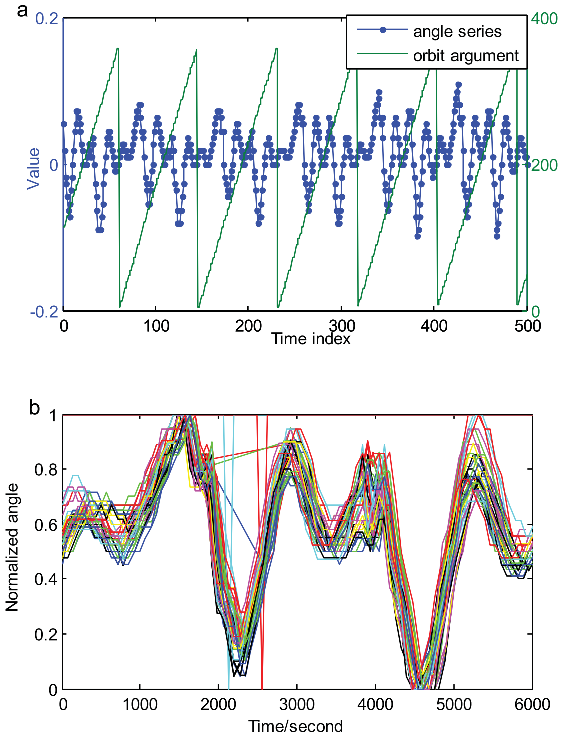

Fig. 9. a) Angle series and orbit argument series. b) Angle subseries based on orbit argument segmentation.

3) For each subseries, SPSegmentation method is used to realize time series representation to improve the efficiency of clustering algorithm.
4) Perform the isometric treatment on each pair of telemetry subseries to obtain the PSPS series. It is worth mentioning that the starting and ending index of each subseries remain the same before isometric treatment.
5) Set the distance measure function to obtain the similarity matrix. The available measure function referring to Euclidean, Spearman, Correlation, DTW and DDTW, etc.
6) Perform AHC on the PSPS series to realize mode identification of telemetry series. The clusters with meaningful explanation will be used for detecting anomalies to further enhance the system reliability.

4. Experimental results and analysis

Since satellite telemetry data has no actual category labels, this paper firstly uses the open datasets that have the similar characteristics with satellite telemetry data to quantitatively evaluate the algorithm performance. These experiments on the public datasets are performed from the following three aspects:

– Perform AHC on the normalized raw series with the distance functions of Euclidean, Spearman, Correlation, DTW and DDTW;
– Realize AHC based on SPSegmetnation with DTW and DDTW measuring function, which are denoted as SPS_DTW and SPS_DDTW respectively;
– Based on isometric treatment and five distance measures, PSPS_euclidean, PSPS_Spearman, SPS_Correlation, PSPS_DTW and PSPS_DDTW, are applied to calculate the similarity matrix for AHC.

Then, the proposed clustering framework is conducted with the real satellite telemetry sequence to examine its applicability.

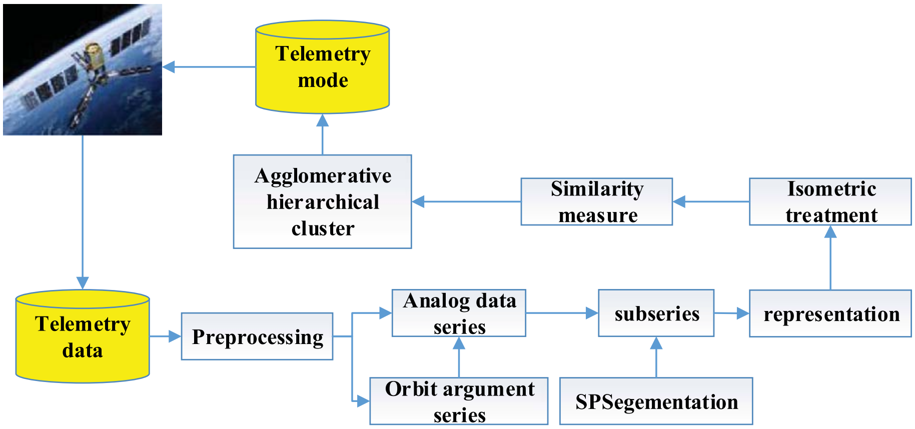

Fig. 10. AHC framework for telemetry series.

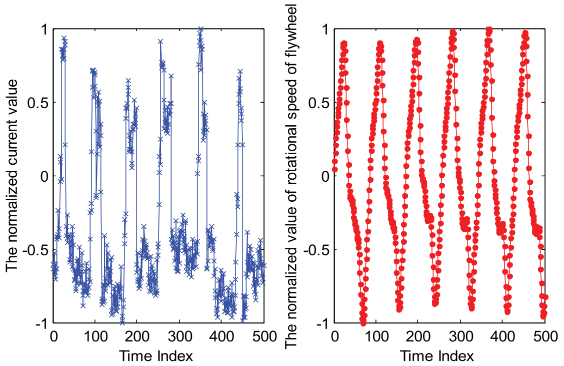

Fig. 11. Two types of satellite telemetry data.

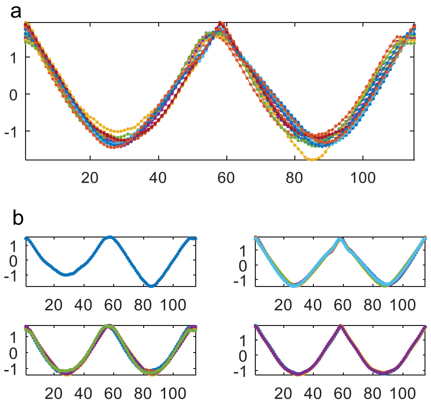

Fig. 12. a) DiatomSizeReduction dataset. b) Four categories of Diatom Size Reduction.

4.1. Data description and evaluation criterion

(1) Data description

Two types of satellite telemetry data, rotational speed of flywheel and current, are shown in Fig. 11 where the values have been normalized and the series have been resampled in about one minute.

Given the shape characteristics of rotational speed and current, the available open source datasets are from the UCR time series database, including the datasets of DiatomSizeReduction and the Face2 which have the similar characteristics with series of rotational speed and current respectively [31]. DiatomSizeReduction dataset includes four categories, and contains 16 series, while each time series has 115 points. The whole dataset and each category of the dataset are shown in Fig. 12.

On the other hand, Face2 dataset is extracted from FaceFour set in UCR database, which includes 2 categories, and each category contains 26 series with 116 points each. Face2 data and each category are shown in Fig. 13.

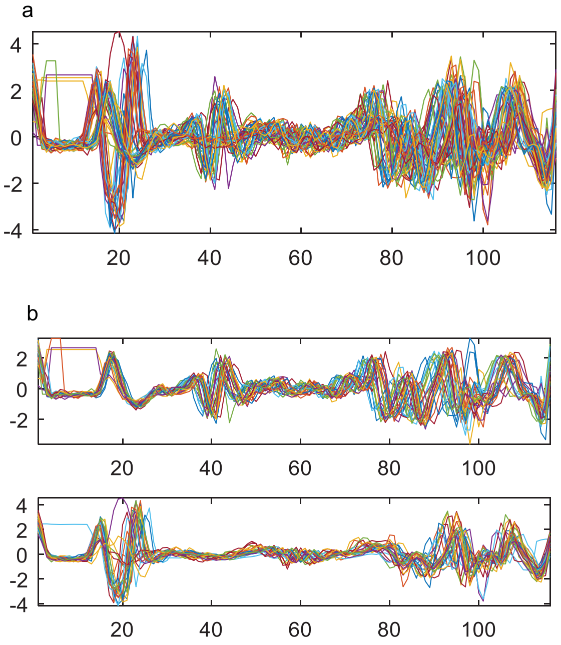

Fig. 13. a) Face2 Dataset. b) Two categories of Face2.

(2) Evaluation criterion

Firstly, Compression ratio, Mean Square Error (MSE) and Mean Absolute Error (MAE) are applied to verify the segmentation performance of feature point's subseries.

Based on the cluster analysis on the open datasets with labels, purity and Adjusted Rand Index (ARI) are the most common external evaluation criterions of cluster accuracy. Bad clustering has lower purity and ARI values close to 0, a perfect clustering has a purity and ARI of 1. The purity can be calculated by Equation (15).

$$purity\,(\Omega, C) = \frac{1}{N}\sum_{K}\max_{J}\left(\left|\omega_k \cap c_j\right|\right) \qquad (15)$$

where $\Omega = \{\omega_1,\ \omega_2, \cdots,\ \omega_K\}$ is the set of clusters and $C = \{c_1,\ c_2, \cdots,\ c_J\}$ is the setting classes. We define ω_k as the kth set of subsequences in Ω and c_j as the jth set of subsequences in C.

The overlap between $\Omega = \{\omega_1,\ \omega_2, \cdots,\ \omega_K\}$ and $C = \{c_1,\ c_2, \cdots,\ c_J\}$ can be merged in a contingency table $[n_{ij}]$.

ARI is defined by Equation (16).

$$AdjustedIndex = \frac{Index - ExpectedIndex}{MaxIndex - ExpectedIndex} \qquad (16)$$

Table 1
The length of SPS and PSPS

C (SPS parameter)	1	...	5	6	...	25
SPS	8~18	8~18	8~18	8~18	7~18	7~18
PSPS	11~26	11~26	11~26	11~26	11~26	11~26

More specifically,

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i\binom{a_i}{2}\sum_j\binom{b_i}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i\binom{a_i}{2} + \sum_j\binom{b_i}{2}\right] - \left[\sum_i\binom{a_i}{2}\sum_j\binom{b_i}{2}\right] / \binom{n}{2}} \tag{17}$$

where $n_{ij} = \left|\omega_k \cap c_j\right|$, a_i, b_j are the values from the contingency table.

4.2. Experiment on DiatomSizeReduction dataset

In order to realize clustering for DiatomSizeReduction dataset based on SPSegmentation, the only parameter needs to be set is the holding time of local extremum C. Noted that the threshold of triangle's midline distance ε can be derived by an adaptive method. Table 1 shows the length of SPS and PSPS with the increasing C.

As is illustrated in Table 1, compared with the raw series with 115 points of each sequence, the amount of data based on SPSegmentation is greatly reduced. Even without the control of C, that is, $C = 1$, SPSegmentation can also get good extraction result which can improve the efficiency of the subsequent clustering algorithm.

Then perform AHC with the following five settings: PSPS_euclidean, PSPS_Spearman, PSPS_Correlation, PSPS_DTW and PSPS_DDTW. Especially, as demonstrated in Section 4.1, in order to make comparison, the clustering is also performed on the normalized raw series with the setting cluster of 4. And the AHC based on SPSegmentation and the distance measure of DTW, DDTW are also performed on the open data set which are denoted as SPS_DTW and SPS_DDTW respectively. Here we give the graphical clustering results on raw data with DTW, SPS_DTW and PSPS_euclidean which are shown in Fig. 14.

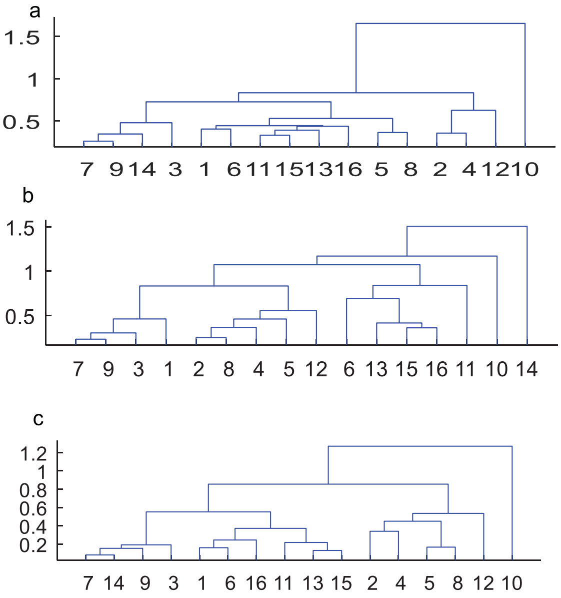

Fig. 14. a) AHC results based on DTW. b) AHC results based on SPS_DTW. c) AHC results based on PSPS_euclidean.

Based on the true labels, the purity and ARI with the setting cluster number of 4 are shown in Table 2.

Table 2
AHC with different inputs and distance measures for DiatomSizeReduction dataset

Inputs_distance measure	Purity	RI
Raw data_Euclidean	1.00	1.00
Raw data_Spearman	0.88	0.82
Raw data_Correlation	1.00	1.00
Raw data_DTW	0.88	0.63
Raw data_DDTW	0.75	0.69
SPS_DTW	0.75	0.42
SPS_DDTW	0.63	0.18
PSPS_Euclidean	1.00	1.00
PSPS_Spearman	0.56	0.12
PSPS_Correlation	0.69	0.50
PSPS_DTW	1.00	1.00
PSPS_DDTW	0.63	0.45

As shown in Table 2, the clustering of raw series with Euclidean and Correlation obtain accurate clustering result. While the clustering on SPSs with DTW or DDTW have worse result compared to the

Table 3
The length of SPS and PSPS

SPS parameter (C)	Length of SPS	Length of PSPS
6	29–39	42–67
7	25–36	34–61
8	22–33	30–57
9	20–32	30–54
10	17–30	25–49
11	16–29	21–48
12	14–27	21–46
13	14–26	20–45
14	13–25	19–42
15	11–25	18–41
16	12–24	16–37
17	11–23	15–39
18	10–25	13–40
19	10–22	16–38
20	10–23	13–38

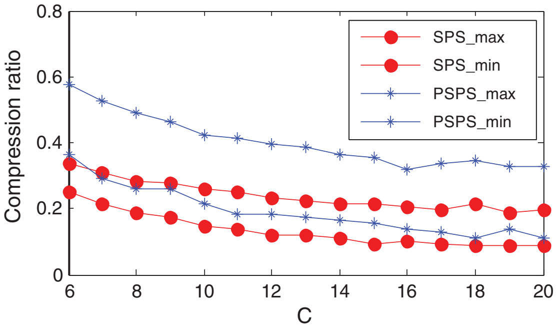

Fig. 15. Change curve of compression ratio.

clustering result on raw data series. Noted that the clustering of PSPS_Euclidean and PSPS_DTW have excellent performance only with 9.6% to 22.6% of the original clustering input which greatly reduce the calculation amount. And the compression ratio is 0.051 and 0.009, respectively.

4.3. Experimental on Face2 dataset

Table 3 shows the length of SPS and PSPS with the increasing C, also the change curve of compression ratio is illustrated in Fig. 15.

As shown in Fig. 15, the length of SPS and PSPS decreases with the increasing of the holding time of extremum.

The average compression ratio of SPS and PSPS are 13.8% and 23.3% (when $C = 19$), respectively, showing that this method effectively reduces the amount of data and greatly improves the speed of calculation.

Table 4
AHC with different inputs and distance measures for Face2 dataset

Inputs_distance measure	Purity	RI
Raw data_Euclidean	0.52	0.00
Raw data_Spearman	0.87	0.53
Raw data_Correlation	0.88	0.58
Raw data_DTW	0.50	0.00
Raw data_DDTW	0.85	0.47
SPS_DTW	0.85	0.47
SPS_DDTW	0.56	0.01
PSPS_Euclidean	0.90	0.65
PSPS_Spearman	0.92	0.71
PSPS_Correlation	0.87	0.53
PSPS_DTW	0.87	0.53
PSPS_DDTW	0.85	0.47

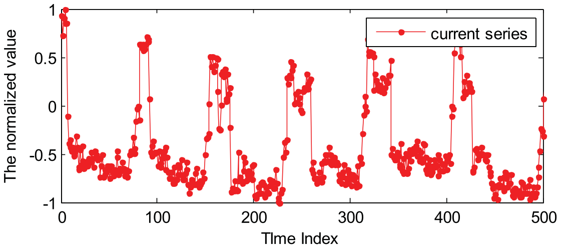

Fig. 16. The normalized current series of satellite telemetry.

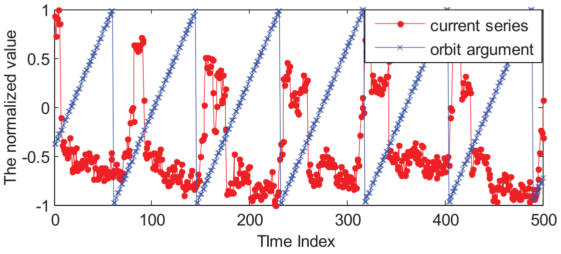

Fig. 17. The normalized telemetry of orbit argument and current.

Similarly, based on the true labels, when the setting cluster number is 2, the purity and ARI of AHC with different inputs and distance measures are shown in Table 4.

Compared with the clustering on original sequence, the clustering input is reduced obviously based on SPSegmentation. At the same time, the quality of the cluster is improved under certain SPS parameters. With the adaptive parameter ε, the average compression ratio of PSPS is 23.3% (when $C = 19$) which will further improve the clustering efficiency.

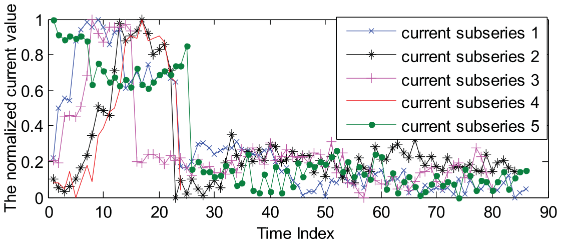

Fig. 18. Current subseries segmented by orbit argument.

4.4. Case study: Clustering on satellite current series

The raw telemetry current with pseudo-period characteristic is shown in Fig. 16 which has been resampled in about one minute. Also, the current values have been normalized.

Then apply the orbit argument, changing from 0 to 360 degrees, to divide the raw current series into some subseries. Figure 17 shows the normalized telemetry of orbit argument and current, the segmentation result of current based on orbit argument is shown in Fig. 18.

In this work, 233 subseries were used in the experiments, and these subseries have different lengths, where the maximus length reaches 104, whereas the shortest subsequence only has 46 points. The mean and median length is 84.5 and 86 respectively. The noise variance is relatively small compared with the trend of current.

Figure 19 shows the change curve of compression ratio, MAE and MSE with the increasing C for current series.

As shown in Fig. 19, the sequence compression ratio gradually decreases with the increasing of C. And the fitting error between the raw series and the

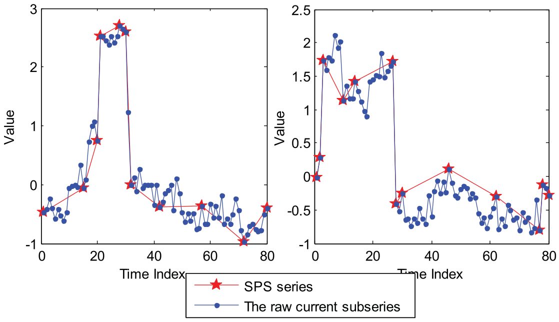

Fig. 20. The SPSs and raw current series of satellite telemetry.

segmentation series is increasing, especially, the sudden enlargement appears at $C = 17$. So, we can select C to be 16 with the acceptable MSE and MAE, where the compression is 12.64%, namely, the 87.36% of raw series can be reduced. Two examples of current SPS series and the corresponding raw current subseries are shown in Fig. 20 where $C = 16$.

As shown in Fig. 20, the feature points represented by SPS can catch the similar shape of the raw series which is great meaningful for the subsequent clustering.

In order to perform AHC on the SPS series with more measure functions, the isometric treatment in pairs should be applied for the SPS series. The length of SPS and PSPS for current telemetry are listed in Table 5.

Due to the true labels of each subsequence are unknown, we only demonstrate the raw subseries with different cluster labels with the setting cluster of 4 and $C = 16$ as shown in Fig. 21, and the measure function is setting to Euclidean distance.

As shown in Fig. 21, these subseries have high intra-cluster similarity, whereas the subsequences in

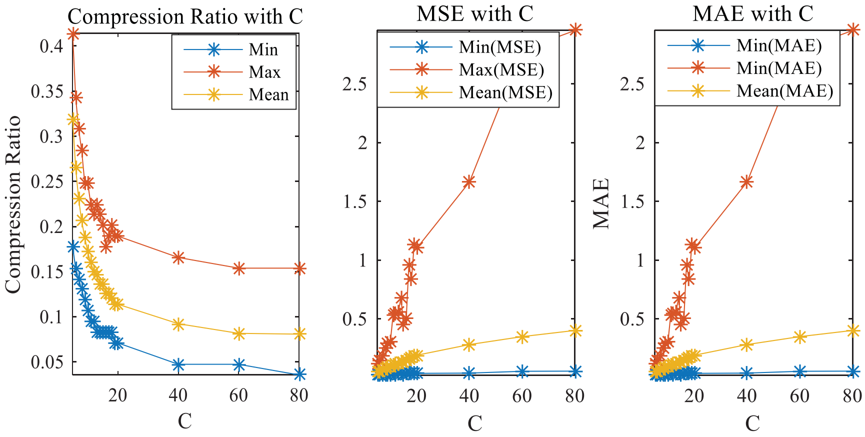

Fig. 19. Compression ratio, MAE and MSE with the increasing C.

Table 5
The length of SPS and PSPS for current telemetry

SPS parameter(C)	Length of SPS	Length of PSPS
6	16–27	23–47
7	15–23	18–40
8	13–22	15–38
9	12–20	14–35
10	11–20	13–34
11	10–19	11–33
12	9–16	11–30
13	8–17	9–30
14	9–17	9–29
15	8–15	9–26
16	8–15	9–27
17	6–15	7–26
18	6–14	7–23
19	6–14	7–25
20	6–15	7–25

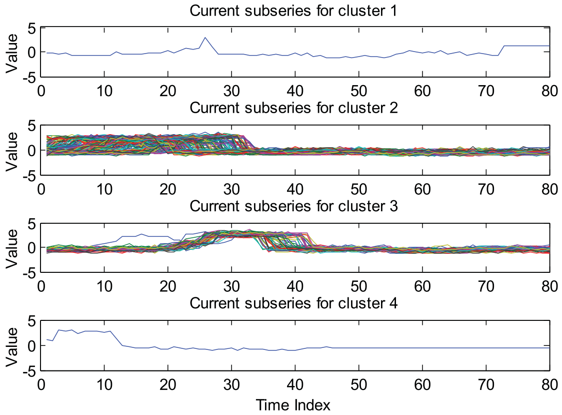

Fig. 21. Clustering result with setting cluster of 4.

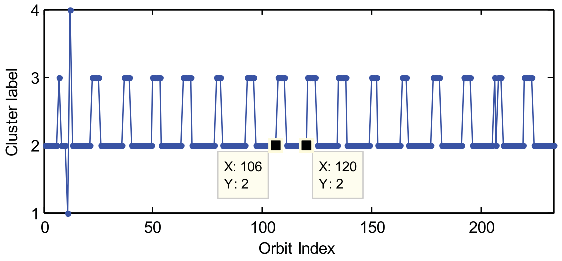

Fig. 22. Clustering labels with the setting cluster of 4.

different clusters have obvious dissimilarity. When we sort the class labels in the order of orbit periods, the result is received as Fig. 22.

As shown in Fig. 22, the clustering labels have strong regularity. In detail, raw current series shows the characteristic of pseudo-period, when it is divided into some subsequences, the clustering labels of these subseries also appear the pseudo-period phenomenon, moreover, most periods of clustering labels include fourteen orbit subseries which lasts about half and five days. Based on the expert knowledge, the satellite orbit is 101.5 minutes, and it goes around the Earth about 14 times, the Quasi repetition time is 5.5 days, so the clustering label changes consistent with the real cases. Hence, with the effective clustering, data visualization can be improved, even some abnormal orbit subseries can be detected with the proposed data clustering framework as shown in Fig. 21.

4.5. *Analysis and discussion*

For the experiments on the datasets of DiatomSizeReduction and Face2, different inputs and distance measures are set as the input of AHC. Given DiatomSizeReduction dataset is smoother than Face2 series, it has higher synchronization feature, the clustering on raw sequence with Euclidean and Correlation has outstanding results. Nevertheless, the clustering with the large reduced input based on PSPS processing can reach the same performance referring to PSPS_euclidean and PSPS_DTW. Especially, the PSPS_Spearman on Face2 has better performance than the clustering on raw sequence.

Therefore, for Isometric Treatment, the accuracy of clustering based on PSPS_Dist is improved compared with SPS_Dist shown in Tables 2 and 4. Note that clustering accuracy is influenced by the data characteristics, distance measures and parameter settings. The clustering based on PSPS can show more improvement on the datasets with some noise and long length sequence.

For the real telemetry series, clustering based on improved time series segmentation can catch the main information of the raw series, moreover, the analysis on these future-points series has a high efficiency which is meaningful for some on-orbit applications., especially, there is no labels for the telemetry data, with the proposed clustering framework, the similar conclusion can be reached with the prior knowledge, so the unknown knowledge may be discovered with the intelligent mining for telemetry series. Furthermore, some abnormal orbit periods can be detected with the improved clustering framework to provide important reference for logistical support.

5. Conclusions

This paper proposes a clustering framework to realize intelligent operating pattern analysis on satellite telemetry data. Firstly, based on the Time-Spatial characteristic analysis of the real telemetry data, the series segmentation is conducted based on the change of orbit argument. Then, improved representation method for time series is designed to extract special point series to improve the clustering efficiency and reduce the noise interference. Once more, many distance measures are introduced to compute the similarity matrix with isometric treatment. Finally, the clustering framework with AHC algorithm is effectively proposed to discover the latent modes within the time series, especially for satellite telemetry series. Both experiments on open source data set and actual monitoring telemetry data set prove the improved efficiency and effective performance.

So far, there are still some works need to be focused in the future. For instance, the number of cluster should be determined by combining the data characteristics and other methods, also more telemetry parameters need to be analyzed based on the clustering framework. In addition, the clustering performance should be further fused with the knowledge-based satellite data analysis.

Acknowledgments

This work was partially supported by National Natural Science Foundation of China under Grant Nos. 61571160, 61771157, 61301205 and the New Direction of Subject Development in Harbin Institute of Technology under Grant No. 01509421.

References

[1] J. Fauste, J. Barreto, M. Casale and D. Ponz, An interactive telemetry data analysis tool for XMM-Newton, *The 9th International Symposium on Space Operations* (*SpaceOps 2006*), Rome, Italy, 2006, pp. 1–8.

[2] T. Yairi, T. Oda, Y. Nakajima, N. Miura and N. Takata, Evaluation testing of learning-based telemetry monitoring and anomaly detection system in SDS-4 operation, *The International Symposium on Artificial Intelligence, Robotics and Automation in Space (i-SAIRAS)*, Montreal, Canada, 2014, pp. 1–10.

[3] R. Santos, How the use of "Big Data" clusters improves off-line data analysis and operations, *International Conference on Space Operations (SpaceOps)*, Pasadena, CA, 2014, pp. 1–8.

[4] J.A.M. Heras, K.L. Yeung, A. Donati, B. Sousa and N. Keil, DrMUST: Automating the anomaly investigation first-cut, *IJCAI-09 Workshop on Artificial Intelligence in Space*, Pasadena, California, USA, 2009, pp. 1–8.

[5] B.R. Mohammad and W.M. Hussein, A novel approach of health monitoring and anomaly detection applied to spacecraft telemetry based on PLSDA multivariate latent technique, *International Workshop on Research and Education in Mechatronics*, El Gouna, Egypt, 2014, pp. 1–6.

[6] A. Barua and K. Khorasani, Hierarchical fault diagnosis and health monitoring in satellites formation flight, *IEEE Transactions on Systems Man* & *Cybernetics* **41**(2) (2011), 223–239.

[7] Y. Song, D. Liu, C. Yang and Y. Peng, Data-driven hybrid remaining useful life estimation approach for spacecraft lithium-ion battery, *Microelectronics Reliability* **75** (2017), 142–153.

[8] I. Beg and T. Rashid, An improved clustering algorithm using fuzzy relation for the performance evaluation of humanistic systems, *International Journal of Intelligent Systems* **29**(12) (2015), 1181–1199.

[9] P. Su, C. Shang and Q. Shen, A hierarchical fuzzy cluster ensemble approach and its application to big data clustering, *Journal Of Intelligent* & *Fuzzy Systems* **28**(6) (2015), 2409–2421.

[10] Q. Miao, X. Zhang, Z.W. Liu and H. Zhang, Condition multi-classification and evaluation of system degradation process using an improved support vector machine, *Microelectronics Reliability* **75** (2017), 223–232.

[11] C. Li, L. Ledo, M. Delgado, M. Cerrada, F. Pacheco, D. Cabrera, R.V. Sánchez and J.V.D. Oliveira, A Bayesian approach to consequent parameter estimation in probabilistic fuzzy systems and its application to bearing fault classification, *Knowledge-Based Systems* **129** (2017), 39–60.

[12] Q. Miao, L. Xie, H. Cui, W. Liang and M. Pecht, Remaining useful life prediction of Lithium-ion battery with unscented particle filter technique, *Microelectronics Reliability* **53**(6) (2013), 805–810.

[13] C. Li, J.V.D. Oliveira, R.V. Sanchez, M. Cerrada, G. Zurita and D. Cabrera, Fuzzy determination of informative frequency band for bearing fault detection, *Journal of Intelligent* & *Fuzzy Systems* **30**(6) (2016), 3513–3525.

[14] J. Pang, D. Liu, Y. Peng and X. Peng, Anomaly detection based on uncertainty fusion for univariate monitoring series, *Measurement* **95** (2017), 280–292.

[15] R. Xu and D. Wunsch, Survey of clustering algorithms, *IEEE Transactions on Neural Networks* **16**(3) (2005), 645–678.

[16] S. Zhou, Z. Xu and F. Liu, Method for determining the optimal number of clusters based on agglomerative hierarchical clustering, *IEEE Transactions on Neural Networks* & *Learning Systems* **99** (2016), 1–11.

[17] O. Santolík, M. Parrot and F. Lefeuvre, Singular value decomposition methods for wave propagation analysis, *Radio Science* **38**(1) (2016), 10-1–10-13.

[18] J. Zhong and Y. Huang, Time-Frequency representation based on an adaptive short-time fourier transform, *IEEE Transactions on Signal Processing* **58**(10) (2010), 5118–5128.

[19] J. Lin, E. Keogh, L. Wei and S. Lonardi, Experiencing SAX: A novel symbolic representation of time series, *Data Mining and Knowledge Discovery* **15**(2) (2007), 107–144.

[20] E.I. Laftchiev, Robust dynamical model-based data representations and structuring of time series data for in-sequence localization, Dissertations & Theses, The Pennsylvania State University, 2015.

[21] Q. Xie, C. Pang, X. Zhou, X. Zhag and K. Deng, Maximum error-bounded piecewise linear representation for online stream approximation, *The VLDB Journal* **23**(6) (2014), 915–937.

[22] C. Guo, H. Li and D. Pan, An improved piecewise aggregate approximation based on statistical features for time series mining, *International Conference on Knowledge Science, Engineering and Management*, Belfast, Northern Ireland, UK, 2010, pp. 234–244.

[23] E. Keogh, S. Chu, D. Hart and M. Pazzani, An online algorithm for segmentingtime series, *Proc of the 1st IEEE International Conference on DataMining*, Washington D.C., USA, 2001, pp. 289–296.

[24] K. Chakrabarti, E. Keogh, S. Mehrotra and M. Pazzani, Locally adaptive dimensionality reduction for indexing large time series databases, *ACM SIGMOD Record* **30**(2) (2001), 151–162.

[25] S. Zolhavarieh, S. Aghabozorgi and Y.W. Teh, A review of subsequence time series clustering, *The Scientific World Journal* **2014** (2014), 1–19.

[26] D. Wang, P.J. Fortier and H.E. Michel, Novel pruning based hierarchical agglomerative clustering for mining outliers in financial time series, *WIT Transactions on Information and Communication Technologies* **41** (2008), 33–42.

[27] S. Park, S.W. Kim and W.W. Chu, Segment-based approach for subsequence searches in sequence databases, *Proceedings of the 2001 ACM symposium on Applied computing*, Las Vegas, Nevada, USA, 2001, pp. 248–252.

[28] S. Sivaranjani, S. Sivakumari and M. Aasha, Crime prediction and forecasting in Tamilnadu using clustering approaches, *International Conference on Emerging Technological Trends (ICETT)*, Kuala Lumpur, Malaysia, 2016, pp. 1–6.

[29] G. Wang, F. Xing, M. Wei, T. Sun and Z. You, Optimization method of star tracker orientation for sun-synchronous orbit based on space light distribution, *Applied Optics* **56**(15) (2017), 4480–4490.

[30] Y. Hu, L. Chen and J. Huang, Space-based pseudo-fixed latitude observation mode based on the characteristics of geosynchronous orbit belt, *Acta Astronautica* **137** (2017), 31–37.

[31] Y. Chen, E. Keogh, B. Hu, N. Begum, A. Bagnall, A. Mueen and G. Batista, The UCR Time Series Classification Archive (2015). URL www.cs.ucr.edu/∼eamonn/time_series_data/