Real-time Satellite Anomaly Data Tagging Based on DAE-LSTM

Caiyuan Xia

School of Computer Science & Engineering

Xi'an Technological University

Xi'an, 710032, China

E-mail: 2969714860@qq.com

Qianshi Yan

School of Computer Science & Engineering

Xi'an Technological University

Xi'an, 710032, China

Abstract—**Spacecraft is the main carrier of human exploration of outer space, exploration and understanding of the Earth and the universe, and the development of spaceflight can promote human civilization andsocial development, and can meet the nee-ds of economic construction, scientific and technological development, security construction, social progress and other aspects. The current global number of satellites in orbit reaches 5,465, of which China has 541. The vigorous development of the space industry symbolizes the steady improvement of the country's comprehensive national power and overall technology. During the operation, the satellite in orbit needs to transmit data to the ground, these data may be subject to interference from various aspects, or even equipment failure, we find these data in real time is very important to reduce losses. The data transmitted by satellite has obvious temporal characteristics, and Long Short-Term Memory (LSTM) network has obvious advantages for processing temporal data, so this paper proposes a BER marking model based on the combination of LSTM network and self-coding technology. By comparing the data before and after noise reduction, a threshold value can be determined, and the BERs can be accurately distinguished by this method. After testing with real satellite temperature data, the accuracy of the model detection reaches a high level.**

Keywords-LSTM Network; Self Coding Technology; Satellite Data; Deep Learning

I. INTRODUCTION

Spacecraft are the main carrier for human beings to explore outer space and explore the understanding of the Earth and the universe. The development of spaceflight can promote human civilization and the development of society, and can meet the needs of economic construction, scientific and technological development, security construction, social progress and other aspects. According to UCS statistics, as of April 30, 2022, the current number of satellites in orbit around the world reached 5465, of which China has 541. The booming development of the space industry symbolizes the steady improvement of the country's comprehensive national power and overall technology.

The satellite will send some collected data or its own parameters to the ground in the normal operation process. The ground receiving station decodes and demodulates the received data and then transmits it to the user side, which then further processes the received data according to the demand. The data received on the ground may be affected by objective factors such as

electromagnetic interference, multipath effect and equipment failure, resulting in some false code points in the data. Very reliable telemetry data are needed for analyzing the operation status of spacecraft in orbit and for various application areas. The study of spacecraft anomaly detection technology can not only detect potential satellite anomalies in time, but also monitor satellite performance, detect satellite faults, locate the root cause of satellite faults, improve the safety of satellite operation in orbit, improve satellite operational life, and improve satellite work efficiency. Telemetry data fault diagnosis and abnormality detection technology has also become an important direction of research in the field of aerospace. The following characteristics exist for its parameter data: the boundary between data in anomalous mode and normal data is blurred and difficult to distinguish; the trend and period of normal data changes continuously as the time and position of the spacecraft in orbit shift; the definition of data anomaly is different in different scenarios and environments. Thus, it can be seen that it is challenging to be able to accurately and effectively represent the possibility of telemetry data as anomalous data.

In-orbit satellites will collect a data every time, which has obvious time-series and real-time. This paper adopts a network structure combining LSTM and self-encoder to mark data anomalies in real time for the purpose of real-time detection of anomalous data.

II. RELATED JOBS

The traditional methods for detecting anomalies in temporal data have been a common research topic, and the traditional methods for detecting anomalies in temporal data are detected by the proximity between the data, and when applying such methods, it is necessary to choose the appropriate proximity. If the distance is too large, it means that the similarity between them is low [1] and can be judged as anomalous. There are also density-based detection methods, the principle is to use the number of anomalous points is much less than the number of normal points, and the distribution of normal data has a certain similarity, so the data of low-density neighborhood is judged as anomalous data [2,3]. Deep learning-based data anomaly detection methods have also been widely proposed in recent years, and literature [4] used a combination of stacked self-coding neural networks (SAE) and long short-term memory neural networks (LSTM) for anomaly detection of process data in industry. The literature [5][6] used noise reduction self-encoder (SDAE) for anomaly detection, but did not consider the problem of real-time detection. The literature [7] uses Markov model, and the data dimension processing in the pre-model stage uses principal component analysis to downscale and human experience to extract features, this method does not make good use of the characteristics of the studied data itself, which leads to limited anomaly detection effect.

The detection method based on manual monitoring combined with thresholds is the most traditional and widely used method in practice, but the occurrence of many anomalies in telemetry data does not cause the value to exceed the limit, and its detection method has disadvantages such as high labor cost and poor scalability of thresholds. The application of anomaly detection methods based on expert systems is considered a more widespread and reliable method, applying expert knowledge (expressed in the form of some rules) to determine anomalies through judgment. However, due to the complex space environment of spacecraft, it is difficult for expert systems to establish perfect and comprehensive rules, and often unknown anomalies occur that cannot be detected by expert systems. Data-driven anomaly

detection method based on machine learning related theory is used to model time series data and identify anomalous patterns that do not conform to normal data through the model output. It is the most intelligent and is currently the most popular method for research on data anomaly detection of spacecraft telemetry data, which is characterized by high data volume and high dimensionality, etc. Since it is targeted at analyzing the change pattern of the data itself and does not rely on expert experience, relatively few of them are currently implemented in ground monitoring systems.

RNN (Recurrent Neural Network) is the most typical model for processing time-series data, and LSTM (Long Short-term Memory), as a variant of RNN, can well solve the gradient disappearance and explosion problems that exist in the training process of RNN. Auto-Encoder (AE) is a kind of neural network that learns the effective representation of sample data by unsupervised learning, and can be used for noise reduction of data. By comparing the data before and after noise reduction of telemetry data, it is easy to find the false code points. In this regard, we construct a network model by combining the two to solve the problem of real-time detection of satellite data outliers.

III. LSTM SELF-CODING THEORY

A. LSTM network

Long and short memory neural network, a special kind of recurrent neural network (RNN), has the problem of gradient disappearance or gradient explosion as time increases in the original recurrent neural network, so it cannot handle data of long sequences. To address the shortcomings of RNN, an improved LSTM model is proposed. Figure 1 shows a single LSTM unit.
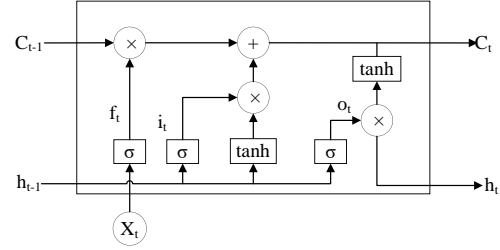
Figer 1.   LSTM unit

In a single LSTM cell, when the input sequence is x_t:

$$f_t = \sigma(W_{hf}h_{t-1} + W_{xf}x_t + b_f) \qquad (1)$$

$$i_t = \sigma(W_{hi}h_{t-1} + W_{xi}x_t + b_i) \qquad (2)$$

$$o_t = \sigma(W_{ho}h_{t-1} + W_{xo}x_t + b_o) \qquad (3)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot tanh(W_{hc}h_{t-1} + W_{xc}x_t + b_c) \qquad (4)$$

$$h_t = o_t \odot tanh(C_t) \qquad (5)$$

The above equation represents the forgetting gate (ft), the input gate (it) and the output of the output gate (ot), respectively, represents the cell state of the current LSTM cell, which will also be used as the input of the next cell, is the output of the current LSTM cell also as the next input. The formula represents the activation function generally uses the sigmoid activation function, W denotes the weight matrix between the forgetting gate and the previous layer, and "$\odot$" denotes the product operation of the elements. From the formula (5), we can see that the connection between the current cell state and the previous cell state is controlled by the forgetting gate, and if the forgetting gate is 0, it means that the previous data is cleared and will not affect the following operation anymore, which also avoids the problem of gradient disappearance or explosion.

B. Self-encoder principle

Self-encoder is a class of artificial neural networks used in semi-supervised and unsupervised learning, whose main function is to extract features and reconstruct data. A complete self-encoding network includes two parts: encoder and decoder, the role of the encoder is to map the input vector into a kind of hidden vector, and the decoder reconstructs the vector into the original input, the network structure is shown in Figure 2. In this paper, a noise reduction self-encoder (DAE) is used, which adds random error points to the input data and learns through training to recover the pure data as much as possible.

Input → Encoder → Hidden vector → Decoder → Output

Error of mean square
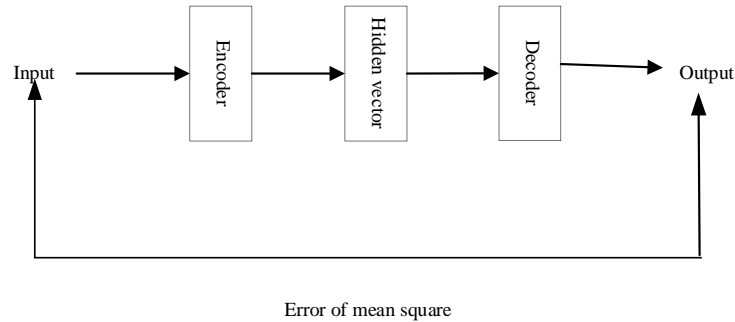
Figer 2.   Self-coding network structure

Equation 6 is the encoding process from the input layer to the hidden layer, where x represents the input, h represents the hidden layer vector, and b is the paranoid term.

$$h = \sigma(W_1 x + b_1) \qquad (6)$$

Decoding process from the hidden layer to the output layer:

$$Output = \sigma(W_2 h + b_2) \qquad (7)$$

The corresponding mean square error loss function is：

$$Loss = \frac{1}{N}(Output - x)^2 \qquad (8)$$

Through iteration after iteration, the training of the self-encoder is completed when the error stabilizes and the input and output are as similar as possible.

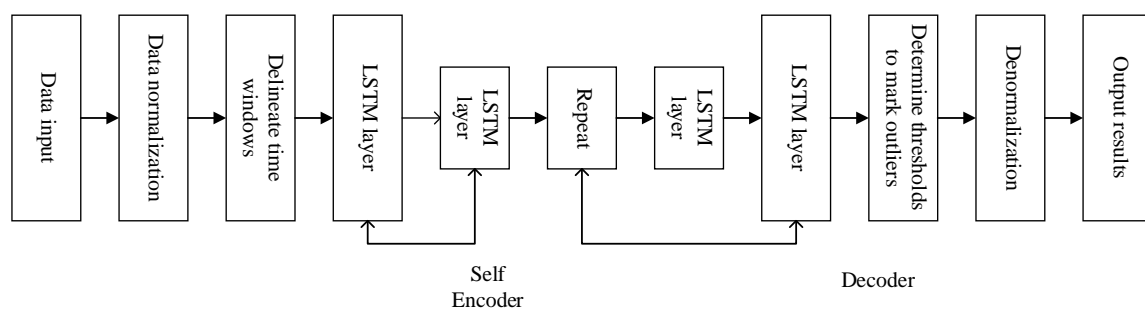IV.   REAL-TIME SATELLITE ANOMALY DATA TAGGING BASED ON DAE-LSTM

A. Pre-processing of data

The data set in the experiment is from a model of satellite taken to the temperature data, in the data obtained, in order to save storage space near the same data was omitted, and in the use of the omitted data need to be restored, the data sampling interval is known 8.192s, so every 8.192s if there is time data missing that copy the previous data as the data value of the moment. The input requirement of the LSTM network is temporal data, so the data needs to be partitioned into multiple sets of time windows, and the length of the time windows also affects the accuracy of the model, for which time windows of different lengths are tested. In order to speed up the convergence of the network, the data are normalized before training, and for comparison during detection, we perform the inverse normalization process after detection, so that it still returns to the range before detection. The

normalization is done using the Z-Score normalization method $Z = \dfrac{x - \bar{x}}{\sigma}$, Retain the variance before normalization σ and the mean value $\bar{x}$ and other data, It is used to denormalize the data after detection, and we can use $x = Z * \sigma + \bar{x}$ to recover the data after denormalization. As the data we got is better quality data, there is no error code point in the data, we need to add error code point manually, also for the normal operation of the evaluation mechanism we save the position of the added error code point, in order to simulate the error code point of the satellite data as much as possible, we add the error code position is random, the number of error code point is set to 1/1000 of the total number of data with reference to the actual situation, considering the short time interval of the data collection of this type of on-board sensor, there may be continuous error code situation, we choose 1/5 of the error code point set as continuous error code point.

B. Experimental process analysis

1) Construction of the network model

By deepening the number of network layers of the self-encoder, more interference information can be suppressed and more effective data feature representation can be obtained. The encoder part of the experiment is a three-layer network structure, and the decoder part is also a three-layer network structure, which is built by using the keras. layers. LSTM network structure in TensorFlow, and the sequential approach is used to build the LSTM network structure in TensorFlow, using sequential approach to build the network model, the number of encoder and decoder is 64, 128, 128, 64, 128, 128. In order to prevent overfitting, a Dropout layer is added to the network model to randomly deactivate some neurons, and the Dropout ratio is set to 0.3. tanh is used as the activation function, mean square error (mse) is used as the loss function, and Adam optimizer is used to adjust the network to reach the optimum, the number of batch training samples is 32, and the number of experimental training rounds (The number of batch training samples is 32, and the number of training rounds (Epoch) is set to 15 for the best results. The overall structure of the program is shown in Figure 3.

Figer 3.   Program structure diagram

2) Number of neurons in the hidden layer

In the physical architecture of the LSTM, an LSTM cell contains four gate structures, and the implementation of each gate structure relies on multiple hidden function layer neurons, which are equivalent to four hidden layers, each of which is fully connected to the input, and the input is composed of two parts, one is the output of the LSTM cell at the previous moment, and another part is the input of the current sample vector, so

that by the After the internal computation of the LSTM unit, the final output of the current moment is obtained as part of the input of the LSTM unit in the next moment. The number of neurons in the hidden layer also has an important impact on the effect of the model. The more the number of neurons, the more complex the network and the higher the accuracy, but it also brings the problem of more computation. From the figure, we can see that the accuracy of the model is better when the number of neurons is 64 or more, and there is no need to increase the number of neurons to avoid the computational overload.
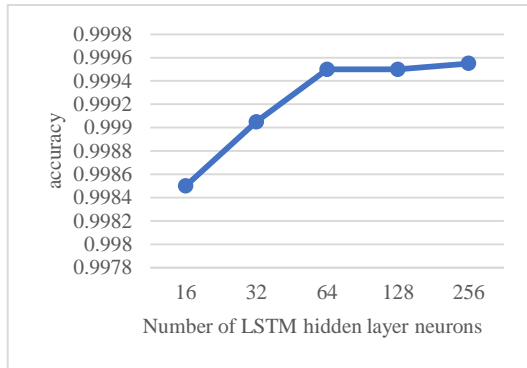

Figer 4.   Effect of the number of neurons on the accuracy of the model

3) Determination of the number of network layers

The number of network layers of the self-encoder largely affects the effect of the model. The deeper the layers, the more theoretically the ability to fit the function is enhanced and the effect will be better, but in practice deeper layers may bring about overfitting problems and also increase the training difficulty, making it difficult to converge the model. Shallow neural networks do not have a high degree of feature abstraction, for which we need to determine the optimal network depth. As shown in Figure 5, the horizontal coordinates indicate the number of network layers for the encoder and the same number of layers for the decoder.
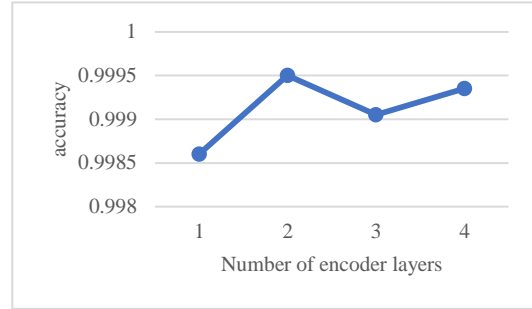

Figer 5.   Effect of the number of network layers on the accuracy of the model

4) Model evaluation mechanism

Regarding how to measure the accuracy in the experiment, we used two metrics to measure it, which are the ratio of the number of abnormal data correctly classified as abnormal to the number of detected abnormal data (acc_n) and the ratio of the number of normal data and abnormal data both correctly classified to the total number of detections (acc). The details are calculated as follows.

$$TP = set(x\_l) \,\&\, set(x) \qquad (9)$$

$$TN = \big(set(testx) - set(x\_l)\big) \,\&\, \big(set(testx) - set(x)\big) \quad (10)$$

$$acc = \big(len(TP) + len(TN)\big) \,/\, len(testx) \qquad (11)$$

$$acc\_n = len(TP) \,/\, len(x\_l) \qquad (12)$$

TP in equation (9) represents the anomalies correctly classified as abnormal, set(x_l) represents the anomalies added to the data during preprocessing, and we read them again during testing, set(x) represents the anomalies found by the model, and the two intersect to get the anomalies correctly found by our model.

TN in equation (10) represents the number of normal data correctly classified as normal, and set (testx) represents the test set data.

The acc in equation. (11) represents the accuracy of the model in correctly classifying abnormal data and normal data.

The acc_n in equation (12) represents the accuracy of the model in correctly classifying the anomalies as abnormal.

The accuracy rates used in the following experiments are all calculated by acc，In addition to measuring by accuracy, we also need to determine whether the network reaches optimality in the training process through the loss function, which determines the number of iterations. We choose the mse (mean squared error) loss function, which uses the difference between the data before and after training squared before taking the mean value.

5) Construction of time series

LSTM networks require the input data to be time-series data, given a set of multivariate time series of length n $X =\{ X_1, X_2,..., X_n \}$, where the observations at the moment t($1 \leq t \leq n$) are m-dimensional vectors $X_t =[x_1, x_2,..., x_m], X$ It records the system state information at this time. The goal of time series data anomaly detection is to determine if there is an anomaly in the time series, and if so, to locate the anomaly. In order to improve the robustness of the detection model, a sliding time window of length k is used to segment the temporal data, and then obtain the set of subsequences $W = \{ W_1, W_2, . . ., W_n \}$, where the subsequence $W_t = \{ X_{t-k+1}, X_{t-k+2},..., X_t \}$ is the window data at moment. Anomaly detection for

data X_t can be translated into anomaly detection for subsequence W_t. Anomaly detection on the data can be translated into anomaly detection on the subseries. When we divided the data into time windows, we experimentally confirmed that different lengths of time windows have an impact on the accuracy of the model. As shown in Figure 6, in order to test the effect of the length of the time window on the accuracy rate, 10% of the error points are randomly added to the data, and it can be seen that the accuracy rate reaches a maximum of 0.995 when the length of the time window is around 15, so the length k of the time window is set to 15 in future experiments.
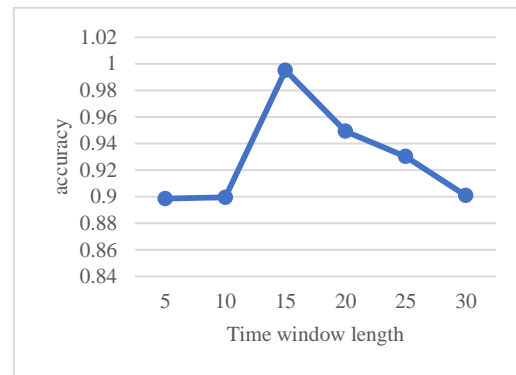

Figer 6.   Effect of time window length on accuracy rate

6) Setting of the threshold value

In the experiment, we use the self-encoder for noise reduction of the data, and the bit error point will be close to the real data, which has little change. Using this feature of the model, we perform the difference operation on the data before and after training, and the difference value exceeds the threshold value we mark as the bit error point. To determine the threshold value, we counted the difference values of the training data and found that most of the differences were within the range of less than 0.25, as shown in Figure 7, for which we set the threshold value to 0.25. The

points exceeding this value were marked as false code points for output.
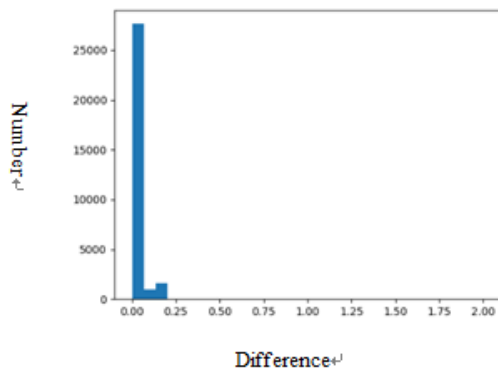

Figer 7.　　Difference statistics

7) Real-time testing

The application background of the model is different from other time-series data processing because satellite data is constantly sent to the ground, the amount of data generated is very large, we need not only the detection of historical data but also the real-time detection of data, if we use a period of data for testing will cause a certain degree of delay for our practical applications, we can not put all the data into the We cannot put all the data into the model for training and testing. For this reason, we adopt a training and testing approach based on the idea of a sliding window, in which a model is trained and saved before starting to detect the data, and when a certain amount of data is detected, training is started again, and the latest saved model replaces the previous one, and the model is updated once for the next data detection. When saving the model, we use the method of saving the network parameters and the network structure separately, as shown in Figure 8, which also reduces the requirement for hardware equipment. We also conducted special experiments on how much data to take for training each time to make the accuracy reach a high level, as shown in Fig. 9, the accuracy rate tends to stabilize after the training amount reaches 50,000,

and the accuracy rate reaches about 99.6%, so we use 50,000 data for training in the experiment. One data point is tested at a time, and the network weights are updated once after the cumulative 50,000 data points are trained. A printing operation is performed after each test of 10,000 data, and the experimental results are printed out as shown in Figure 10.
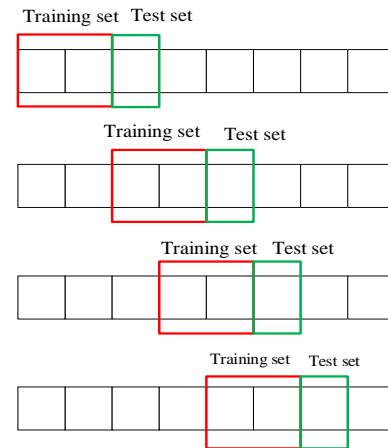

Figer 8.　　Program execution proces
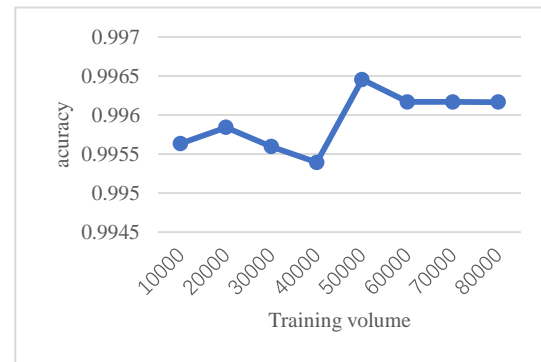
Figer 9.　　Relationship between training volume and accuracy rate

C. Analysis of experimental results

In the experiment, 50,000 data are used as training, one data is tested each time, and the data are iterated 15 times. As shown in Figure 10, the first image is the image before data processing, the second image is the image after noise reduction using the trained model, and the third image is the identified error code points. After the actual test, the accuracy of the model reaches

99.95%，It can be considered that our model achieves a better result, after which we would like to add other types of satellite data for comparative analysis.
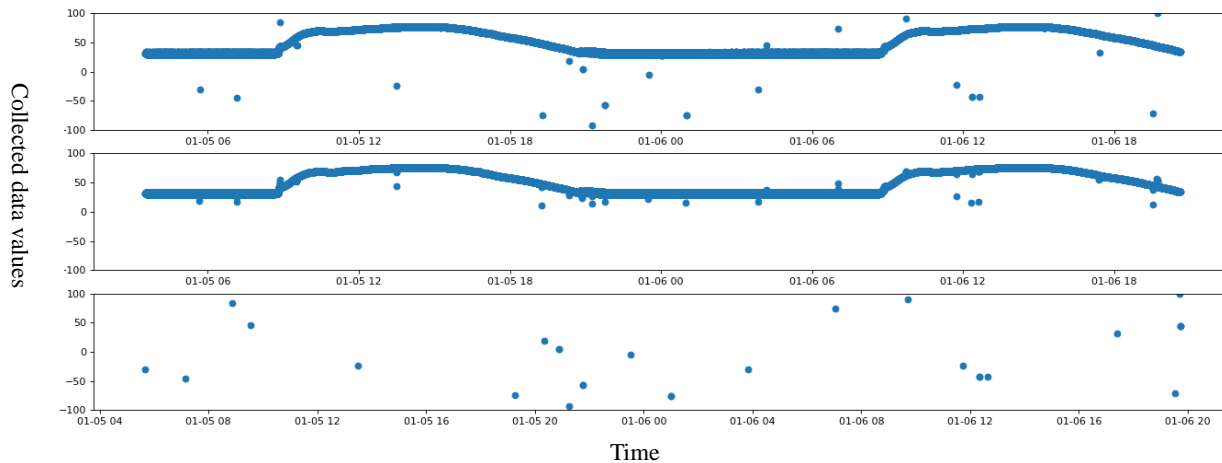
Figer 10.   Experimental results

V.   CONCLUSION

Spacecraft are multi-functional and complex systems that can be used in many fields, such as navigation satellites and meteorological satellites, etc. Therefore, spacecraft are high-value targets, and it is of great importance to ensure their safe and stable operation. The only means to obtain the operational status of spacecraft is telemetry data, which can reflect the orbital operation of spacecraft and the operational status of each component. Due to the communication or sampling reasons, some of the telemetry data may contain wrong data (abnormal data), which can greatly interfere with the subsequent spacecraft analysis and even make the fault detection and health analysis of the spacecraft impossible. Therefore, certain means are needed to analyze the telemetry data to evaluate the accuracy of the data itself and prepare for further analysis of the spacecraft. However, there is no effective means to analyze the reliability of telemetry data. Therefore, by studying the characteristics of telemetry time series data and the existing time series data analysis methods, this paper focuses on the anomaly detection method of telemetry time series data with spacecraft telemetry data as the research object and uses the self-encoder related technology.

REFERENCES

[1] Kullback S., Leibler R. A..   On information and sufficiency [J]. The Annals of Mathematical Statistics, 1951, 22(1): 79-86

[2] Zhang L. W., Lin J., Karim R.. Adaptive kernel density-based anomaly detection for nonlinear systems [J]. Knowledge-Based Systems, 2018, 139: 50-63

[3] Liu F, Qi JP, Yu YW, et al. Density-based top-n local anomaly fast detection algorithm[J]. Journal of Automation, 2019, 45(9): 1756-1771

[4] Shang W-L, Shi He, Zhao J-M, et al. An SAE-LSTM-based method for process data anomaly detection [J]. Journal of Electronics, 2021, 49(08): 1561-1568.

[5] Liu Yongmei. Research on the anomaly detection method of satellite power parameters based on SDAE [D]. Harbin Institute of Technology, 2021.

[6] Guo Lei. Research on rolling bearing fault diagnosis based on SDAE and LSTM fusion [D]. Shenyang University of Architecture, 2020.

[7] Zhang, R. L., Wang, W., Guan, X. H.. Hidden Markov model-based anomaly detection for program behavior [J]. Journal of Xi'an Jiaotong University, 2005, 39(10): 1056-1059.

[8] Cuan-Ying,Wu-Yue. Research on time series anomaly detection method based on Seq2Seq depth self-encoder [J]. Modern Electronics Technology, 2022, 45(02): 26-30.

[9] Jin Yupeng. Research on key issues of air pollutant time-series prediction based on improved self-coding network [D]. Shanghai Normal University,2019.

[10] DU Xiaolei,CHEN Zhigang,XU Xu,et al. Research on bearing fault diagnosis based on wavelet convolutional self-encoder and LSTM network [J]. Mechatronics Engineering, 2019, 36(07): 663-668.

[11] Cao Lichang, Deng Dexin, Shen Jian, et al, Meng Haitao. Research on integrated monitoring technology for monitoring station data quality [C]. Proceedings of the 12th Annual China Satellite Navigation Conference - S06 Time Reference and Precision Timing. 2021:54-58.

[12] Chen, J.S., Li, J., Chen, W.G., et al. Wind turbine state anomaly detection method using sliding window and multiple additive noise ratio stacking noise reduction self-coding [J]. Journal of Electrical Engineering Technology, 2020, 35(02): 346-358.

[13] Chen J, Liu M, Xiong P, et al. Noise reduction of ECG signals based on convolutional self-coding neural networks [J]. Computer Engineering and Applications, 2020, 56(16): 148-155.

[14] Yuanyuan Wei, Julian Jang-Jaccard, et al. LSTM-Autoencoder based Anomaly Detection for Indoor Air Quality Time Series Data [J]. arXiv preprint arXiv: 2204. 06701v1, 2022.

[15] Victoria Bell, Divish Rengasamy, Benjamin Rothwell , Grazziela P Figueredo. Anomaly Detection for Unmanned Aerial Vehicle Sensor Data Using a Stacked Recurrent Autoencoder Method with Dynamic Thresholding [J]. arXiv preprint arXiv: 2203. 04734, 2022.