Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 4, pp. 2540–2547, DOI 10.12694/scpe.v25i4.2832

COMPLEX EVENT INFORMATION MINING AND PROCESSING FOR MASSIVE AEROSPACE BIG DATA

LIN LI[*]AND LIBIN JIA[†]

Abstract. This paper intends to analyze the existing problems of remote sensing data from the perspectives of space remote sensing information data capacity and data types. Then, a framework for rapidly analyzing and processing space remote sensing information is constructed. Then, LSTM is used to realize the fault diagnosis of remote sensing data continuity, discrete sample mixing and strong correlation of sample variation. LSTM conducts a multimodal analysis of remote-control commands, which is applied to modeling. The multi-stage LSTM prediction model is established and integrated efficiently to improve its adaptive ability in complex space environments. In this way, the anomaly recognition of remote sensing information is realized. Experiments show that the algorithm can improve the anomaly detection rate of remote sensing data. Experiments show that the algorithm is feasible. It can provide reliable data interpretation function for space remote sensing information control system.

Key words: Space remote sensing; Big data; Integrated extended short-term memory network; Hadoop; Map reduce

1. Introduction. With the development and function of spacecraft, its development cycle is shorter and shorter, and the number of launches is increasing. Simply relying on conventional reliability engineering technology has been unable to meet the needs of data processing and analysis. Most of the current satellites are realized by ground-based telemetry. While working in orbit, the ground tone center needs to collect and process the remote sensing data of the satellite in orbit. In this way, the automatic judgment and early warning of the status monitoring parameters of the orbiting satellite can be realized. The obtained remote sensing data are saved in the database for subsequent analysis. The detection of anomalous phenomena in satellite remote sensing signals is one of the research hotspots in the world. The existing remote sensing monitoring technology can be summarized into three categories: statistical direction, distance direction and error direction. Most traditional statistical studies assume that the sample satisfies a specific distribution pattern. The data that does not conform to the distribution or the statistical characteristics that are not consistent are labeled as anomalies. And these samples are often difficult to describe by pre-set distributions. At this point, the relative positions between multiple samples are measured, and the samples close to them are found and labeled. Literature [1] uses density-based, supplemented by ranging technology for high-precision time series anomaly detection. Literature [2] studies the similarity measure based on the temporal features of remote sensing to eliminate the influence of correlation among parameters. This enables non-synchronous measurement.

An extended short-term memory network (LSTM) is a typical recurrent neural network. This method can infer and predict based on historical data while maintaining the advantages of traditional recurrent neural networks. Therefore, the anomaly recognition of time series data has become the focus of current research. In reference [3], National Aeronautics and Space Administration (NASA) integrated nonlinear modeling and automatic extraction characteristics of LSTM and constructed the LSTM model with remote control commands and remote sensing information as inputs. In this way, remote sensing information can be labeled efficiently.

However, the existing LSTM modeling methods have some problems, such as high overall prediction deviation and more variation, and their ability to identify situation and aggregation anomalies needs to be improved [4]. Therefore, this project intends to study the new satellite remote sensing data anomaly detection technology based on LSTM. The clustering idea is used to explore the differences among high dimensional remote-control

[*]Zhengzhou University of Aeronautics, Zhengzhou, Henan 450000, China
[†]Zhengzhou University of Aeronautics, Zhengzhou, Henan 450000, China (jialb190992@163.com)

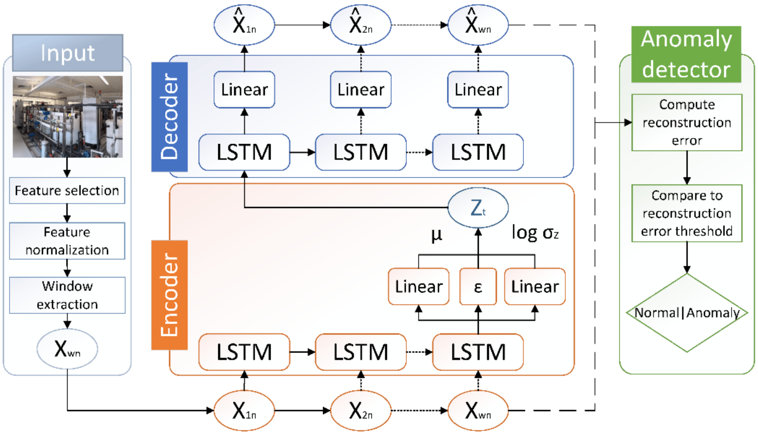

Fig. 2.1: *Anomaly detection method of telemetry data with integrated LSTM prediction model.*

commands efficiently. The LSTM prediction model is constructed for each control mode, and the prediction error threshold is integrated to improve the prediction performance of remote sensing data.

2. Using LSTM integrated forecasting technology to detect space remote sensing data anomalies . When the spacecraft is working in orbit, the difference in the mode characteristics of the telemetry signal under the action of multiple remote-control commands is not considered [5]. This affects the ability to detect remote sensing parameters. This paper will combine LSTM for data processing and divide the workflow into three levels: 1) mining control commands, 2) Model training of multiple LSTMS, and 3) Telemetry data anomaly detection model. The anomaly detection method flow of remote sensing data integrated with the LSTM prediction model is shown in Figure 2.1 (the picture is quoted in Sensors 2022, 22(8), 2886).

2.1. Control guide pattern discovery. First, the algorithm preprocesses the learning process of LSTM. Then, it is classified by clustering based on the time sequence segmentation of fixed units [6]. In this way, the preprocessing of training samples and the discovery of the control mode of remote operation are realized.

2.1.1. Training data preprocessing. The training set D at time t can be viewed as the set of m dimensional vectors, $d^t = \{d_1^t, d_2^t, \mathrm{L}, d_m^t\}$. Here $\left\{d_1^t, d_2^t, \mathrm{L}, d_{m-1}^t\right\}$ is the multi-dimensional remote-control indication for time t. Let $\{d_m^t\}$ be the remote sensing data at time t, then the remote sensing data training set D containing a subsystem is represented as follows

$$D = \left\{ \begin{bmatrix} d_1^1 \\ \mathrm{M} \\ d_{m-1}^1 \\ d_m^1 \end{bmatrix}, \begin{bmatrix} d_1^2 \\ \mathrm{M} \\ d_{m-1}^2 \\ d_m^2 \end{bmatrix}, \mathrm{L}, \begin{bmatrix} d_1^t \\ \mathrm{M} \\ d_{m-1}^t \\ d_m^t \end{bmatrix}, \mathrm{L} \right\}$$

Then the paper reconstructs $D.B'$ is decomposed into multiple submatrices D_n^* containing continuous-time vectors [7]. Each submatrix contains c_t times of telemetry data and global remote-control commands, whose reconstructed input D' is expressed as:

$$\mathcal{D}^0 = \{\{D_1^*\}, \{D_2^*\}, \mathrm{L}, \{D_n^*\}, \mathrm{L}\}$$

Among them:

$$D_n^* = \left\{ \begin{bmatrix} d_1^n \\ d_2^n \\ \mathrm{M} \\ d_m^n \end{bmatrix}, \begin{bmatrix} d_1^{n+2} \\ d_2^{n+2} \\ \mathrm{M} \\ d_m^{n+2} \end{bmatrix}, \mathrm{L}, \begin{bmatrix} d_1^{c_i+n} \\ d_2^{c_i+n} \\ \mathrm{M} \\ d_m^{c_i+n} \end{bmatrix} \right\}$$

2.1.2. Excavation of remote-control mode. A spacecraft remote control method based on real-time monitoring is proposed to solve the incoordination problem in spacecraft remote control [8]. The total data in the sub-training set should not be too small to avoid overfitting and underfitting. Two examples of cluster classes are given, and the clustering method according to the dominant mode is illustrated. For example, remote command R_n^* is extracted from each submatrix D_n^* of B' to obtain the corresponding control command matrix R^* of D' :

$$R_n^* = \left\{ \begin{bmatrix} d_1^n \\ d_2^n \\ \mathrm{M} \\ d_{m-1}^n \end{bmatrix}, \begin{bmatrix} d_1^{n+1} \\ d_2^{n+1} \\ \mathrm{M} \\ d_{m-1}^{n+1} \end{bmatrix}, \mathrm{L}, \begin{bmatrix} d_1^{c_1+n} \\ d_2^{c_2+n} \\ \mathrm{M} \\ d_{m-1}^{c_i+n} \end{bmatrix} \right\}$$

R^* series of control mode characteristic vector $\|R^*\|_2$ is obtained by $C2$ norm operation on a submatrix.

$$\|R^*\|_2 = \{\|R_1^*\|_2 , \|R_2^*\|_2 ,\ \mathrm{L}, \|R_n^*\|_2 ,\ \mathrm{L}\}$$

$$\|R_n^*\|_2 = \sqrt{\eta_{\max}\left(R_n^{*T} R_n^*\right)}$$

Reorder the numeric values in the control graphics feature vector $\|R^*\|_2$. Taking the middle value of $\|R^*\|_2$ as the threshold, the control command set is divided into two categories in the order similar to the control commands.

2.2. Multi-stage LSTM model training. The modal analysis of remote-control commands is carried out, and they are grouped into two sub-training groups. LSTM predicts it. The reconstructed training matrix D is used as input in the forecast of telemetry data [9]. The telemetry data for the next point in time is forecast based on the telemetry data for c_t times contained in each submatrix D_n^* and the corresponding global remote-control command. Assuming $t = c_t$, its prediction steps are:

$$D_n^* = \left\{ \begin{bmatrix} d_1^n \\ d_2^n \\ \mathrm{M} \\ d_m^n \end{bmatrix}, \mathrm{L}, \begin{bmatrix} d_1^{n+t} \\ d_2^{n+t} \\ \mathrm{M} \\ d_m^{n+t} \end{bmatrix} \right\} \to \begin{bmatrix} d_2^{n+t+1} \\ d_2^{n+t+1} \\ \mathrm{M} \\ \hat{f}^{n+t+1} \end{bmatrix}$$

$\hat{f}^{n+t+1}$ is the predicted value at the point $n+t+1$ in time. In constructing the prediction model, the absolute error function (MAE) is the loss function of the LSTM prediction model:

$$\mathrm{MAE}(f, \hat{f}) = \frac{\sum_{i=1}^{n} \left| f_i - \hat{f}_i \right|}{n}$$

n is the total number of all expected values. f_i is the i actual value. This project intends to adopt the LSTM prediction method based on Adam to enhance the sensitivity and accuracy of different control modes through repeated learning [10]. This article sets a specific number of memory modules in each hidden layer as part of the LSTM model to avoid overmatching. LSTM prediction models α and β were obtained after all the sub-training sets were trained.

2.3. Anomaly detection mode of remote sensing data. LSTM was used to train α and β prediction models to obtain two different sequences. Then, the two prediction sequences are weighted and integrated to form the final prediction sequence [11]. The test data is reassembled using a training set B' as an example. LSTM was used to recombine the remote sensing data of α and β, and two sets of remote sensing data prediction sequences $\alpha, \hat{f}_\alpha = \left\{ \hat{f}_\alpha^1, \hat{f}_\alpha^2,\ \mathrm{L}, \hat{f}_\alpha^t,\ \mathrm{L} \right\}$ and $\beta, \hat{f}_\beta = \left\{ \hat{f}_\beta^1, \hat{f}_\beta^2,\ \mathrm{L}, \hat{f}_\beta^t,\ \mathrm{L} \right\}$ were obtained. Finally, the weighted matrix $\hat{F}$ is used to integrate the models and obtain the final forecast sequence.

$$\hat{f} = \left\{ \begin{matrix} \hat{f}_\alpha^1 & \hat{f}_\alpha^2 & \cdots & \hat{f}_\alpha^t & \cdots \\ \hat{f}_\beta^1 & \hat{f}_\beta^2 & \cdots & \hat{f}_\beta^t & \cdots \end{matrix} \right\}$$

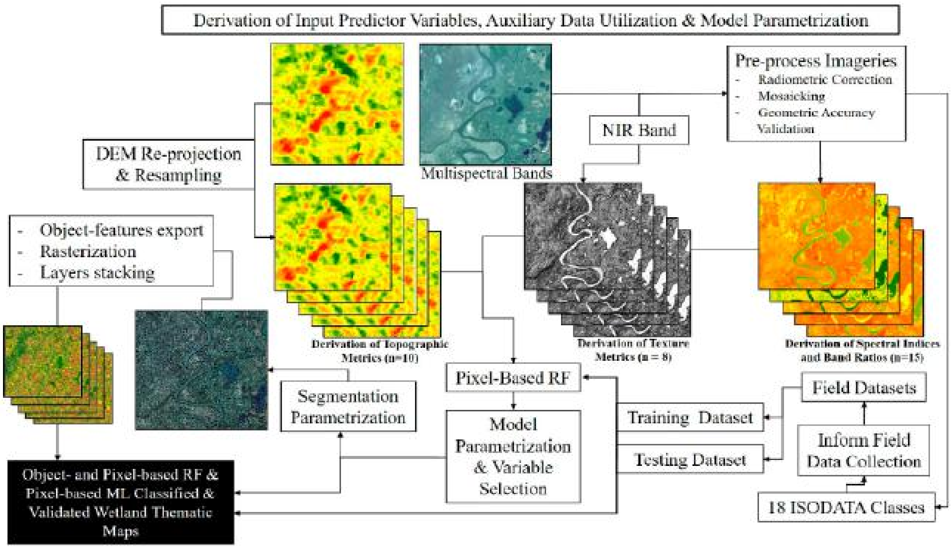
Derivation of Input Predictor Variables, Auxiliary Data Utilization & Model Parametrization
DEM Re-projection & Resampling
Multispectral Bands
NIR Band
Pre-process Imageries
- Radiometric Correction
- Mosaicking
- Geometric Accuracy Validation
- Object-features export
- Rasterization
- Layers stacking
Derivation of Topographic Metrics (n=10)
Derivation of Texture Metrics (n = 8)
Derivation of Spectral Indices and Band Ratios (n=15)
Pixel-Based RF
Segmentation Parametrization
Model Parametrization & Variable Selection
Training Dataset
Testing Dataset
Field Datasets
Inform Field Data Collection
18 ISODATA Classes
Object- and Pixel-based RF & Pixel-based ML Classified & Validated Wetland Thematic Maps

Fig. 3.1: *Remote sensing data processing flow.*

D_t^* is the input to time t, and the corresponding remote control command matrix R_t^* obtains the prediction value $\hat{f}_\alpha^{t+c_t+1}$ of time $t+c_t+1$ through the prediction mode D. The prediction value $\hat{f}_\beta^{t+c_t+1}$ is also obtained from the prediction model β of R_t^* to $t+c_t+1$. If the corresponding norm $C2\left\|R^*\right\|_2$ of R_t^* is the training set α, then:

$$\hat{f}^{t+c_t+1}=\varphi^{t+c_t+1}\hat{f}_\alpha^{t+c_t+1}+\left(1-\varphi^{t+c_t+1}\right)\hat{f}_\beta^{t+c_t+1}$$

If R_t^* corresponds to $C2$ norm $\left\|R^*\right\|_2$ belongs to training set β :

$$\hat{f}^{t+c_t+1}=\left(1-\varphi^{t+c_t+1}\right)\hat{f}_\alpha^{t+c_t+1}+\varphi^{t+c_t+1}\hat{f}_\beta^{t+c_t+1}$$

Gets the remote-control command matrix associated with the current forecast. Dynamically adjust the weights. Finally, A series of prediction sequence $\hat{F}$ is obtained.

3. Space remote sensing cloud processing technology. Remote sensing cloud technology aims to extract new information from massive data. Then, the data is analyzed, reasoned and processed. The Remote sensing data processing process is shown in Figure 3.1 (the image is quoted in Remote sensing, 2017, 10(1): 46). Data preprocessing includes three parts: ”data depth processing,” ”orthographic correction” and ”image fusion and conversion.” Remote sensing data processing is a typical data-intensive computation. This puts forward higher requirements for data processing methods. In the existing cluster system, task division, task processing and network communication are hardware-based. At the same time, data-intensive computation requires freedom from traditional hardware programming patterns. This allows the application to be described using higher-level semantics [12]. Remote sensing cloud technology integrates the relevant technologies in remote sensing data processing into the cloud to realize the storage, processing and transmission of remote sensing data. With the advent of big data, massive remote sensing data has exceeded the existing storage and computing capabilities. Remote sensing cloud computing can provide storage and analysis tools for remote sensing big data. Cloud computing platform administrators use the cloud storage and computing functions provided by the Internet to users. Cloud computing services include Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Software as a Service (SaaS), and business intelligence. Amazon’s EC2 is a type of IaaS. Google and Microsoft are PaaS. IaaS is the cornerstone of future IT industry development. The deployment of network and cloud computing and the storage and processing of massive data will promote the development of IaaS.

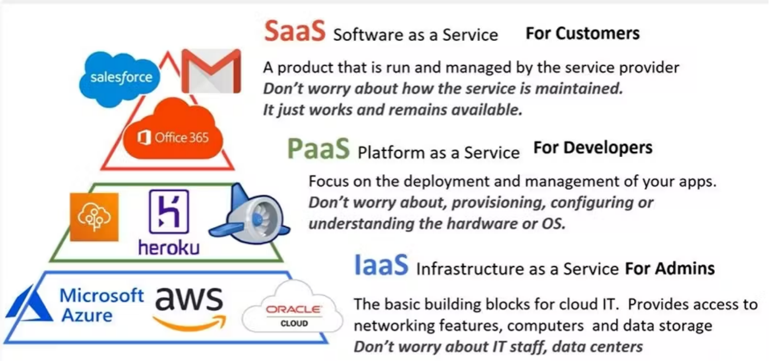
SaaS Software as a Service For Customers
A product that is run and managed by the service provider
Don't worry about how the service is maintained.
It just works and remains available.
PaaS Platform as a Service For Developers
Focus on the deployment and management of your apps.
Don't worry about, provisioning, configuring or understanding the hardware or OS.
IaaS Infrastructure as a Service For Admins
The basic building blocks for cloud IT. Provides access to networking features, computers and data storage
Don't worry about IT staff, data centers
salesforce
Office 365
heroku
Microsoft Azure
aws
ORACLE CLOUD

Fig. 3.2: *Cloud service pyramid model.*

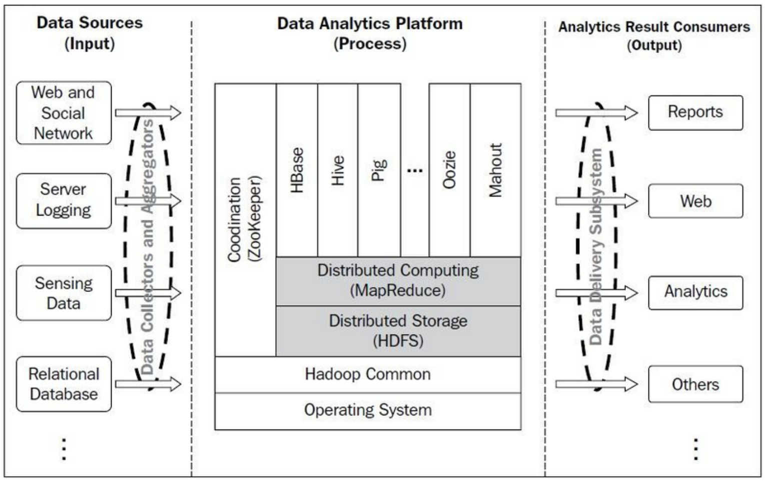

Fig. 3.3: *Hadoop ecosystem architecture.*

PaaS is considered the ”operating system” of the future IT industry. SaaS plays a role in communicating with customers. The three service levels can also be re-split. Each tier contains many kinds of businesses, forming a cone pattern similar to the cloud computing services in Figure 3.2. Hadoop is the most popular large-scale data analysis platform based on the Map Reduce framework. Its core idea is the large-scale, data-intensive, parallel application based on the MapReduce framework. Data storage and processing are realized using the Hadoop distributed file system and Map Reduce programming model. The Hadoop ecosystem is shown in Figure 3.3. The typical cloud model can be divided into three levels: the low-level programming model, the base-valued programming model, and the generic parallel programming model. The underlying programming mode mainly includes OpenMP mode and Message Passing Interface (MPI) mode. Map Reduce, introduced by Google, is a typical Key-Value programming model. It uses the idea of divide and rule. The complex problem is divided into several subproblems by recursion until the subproblem can be solved directly. Standard parallel programming includes standard class libraries such as STL and QUAFF.

Among them, Hadoop-GIS is a platform for processing, retrieving and acquiring massive spatial data. ESRI Hadoop is a tool developed specifically for the Hadoop process. Spatial Hadoop uses Map Reduce mode to construct indexes for building spatial data. TTA REEG uses the Map Reduce system based on Spatial Hadoop, whose core is to process the data in OpenStreetMap. CG Hadoop extends Map Reduce to support polygon, boundary, convex set analysis, farthest point, nearest point and other geometric operations [13]. The space

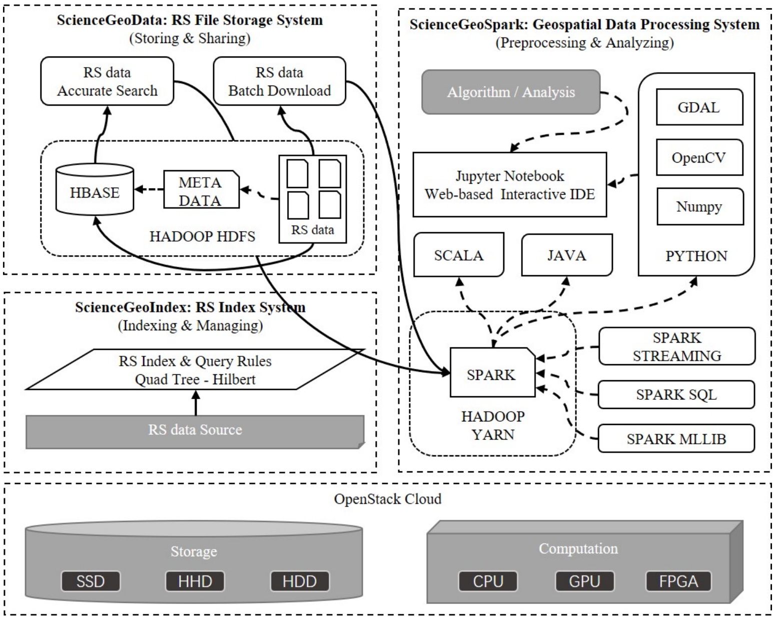

Fig. 3.4: *Remote sensing extensive data processing system based on cloud computing.*

Table 4.1: *Data parameters.*

	SMAP	MSL	Total
Total number of abnormal sequences	72	38	109
Total number of abnormal sequences	66	55	61
Context exception total	40	49	43
Total number of telemetry channels	57	28	49
Total telemetry data detected	447641	69489	517129

remote sensing extensive data processing system based on cloud computing technology is shown in Figure 3.4. The system consists of a remote sensing data acquisition system, a remote sensing data processing system, an extensive data analysis system and a cloud computing system. After the data is preprocessed, the collected data is transmitted to the earth's surface through the downlink, and the remote sensing data processing system is used to store, manage and serve the remote sensing data. Filter data based on availability and distribute processing requests to individual servers using load balancing. After mathematical and logical processing, the result of processing is produced. The paper can focus, edit, analyze, and decide on the data by analyzing the data.

4. System inspection. The confirmation data used in this paper comes from the ISA Report published by NASA, which includes the SMAP and MSL spacecraft models. By comparing with radial neural network (RBF), the detection effect of remote sensing anomaly detection algorithm based on LSTM on point and context is studied.

4.1. Experimental data. The observation data and corresponding remote-control commands of each channel of SMAP and MSL 12 subsystems are used for training and testing. Among these samples, 72 (66.06%) belonged to SMAP and 38 (33.94%) belonged to MSL. Each set contains a different amount of telemetry data and 24 sets of remote commands [14]. The total number of telemetry values for each set is 300-4000. The specific number of training data, the total number of abnormal sequences, the total number of test channels and the total number of test data are shown in Table 4.1.

Table 4.2: *Inspection and comparison of space telemetry anomaly data integrated with LSTM.*

			RBF	Integrated LSTM
SMAP	Point anomaly	Detection rate	88	99
		False alarm rate	10	37
	Context exception	Detection rate	80	92
		False alarm rate	41	56
MSL	Point anomaly	Detection rate	71	71
		False alarm rate	12	12
	Context exception	Detection rate	68	79
		False alarm rate	0	12
Total	Point anomaly	Detection rate	74	81
		False alarm rate	11	34
	Context exception	Detection rate	75	84
		False alarm rate	32	49

4.2. Space telemetry anomaly data inspection integrated with LSTM. Specific test data are shown in Table 4.2. It can be seen from the data in Table 4.2 that compared with the RBF network, the overall anomaly detection rate of the algorithm has increased from 74% to 81%, and the overall anomaly detection rate has increased from 75% to 84%. The algorithm in this paper can make the detection rate of single-point exceptions in SMAP reach more than 95%, while the context exception is stable at about 88%. MSL's single-point anomaly detection rate is about 73%, and the context anomaly is basically stable at about 74%. The single-point anomaly detection rate of Class 2 satellites is about 81%, and the context anomaly detection rate is about 84%. The results show that the point set and the content discovery have greatly improved.

5. Conclusion. A cloud computing platform for space remote sensing information is constructed. A new technique for anomaly detection of LSTM satellite remote sensing data is proposed. The technology is restricted by multiple control modes of remote command, which leads to the degradation of its overall forecasting performance. Multiple LSTM models are established, and the predicted values of each sub-model are weighted and integrated. An exception recognition method based on a dynamic threshold is established. The target can be identified effectively by analyzing the deviation between the data and the actual data.

REFERENCES

[1] Zeadally, S., Siddiqui, F., Baig, Z., & Ibrahim, A. (2020). Smart healthcare: Challenges and potential solutions using Internet of things (IoT) and big data analytics. PSU research review, 4(2), 149-168.

[2] Smys, D. S., Basar, D. A., & Wang, D. H. (2020). CNN based flood management system with IoT sensors and cloud data. Journal of Artificial Intelligence and Capsule Networks, 2(4), 194-200.

[3] Alshammari, H., El-Ghany, S. A., & Shehab, A. (2020). Big IoT healthcare data analytics framework based on fog and cloud computing. Journal of Information Processing Systems, 16(6), 1238-1249.

[4] Raj, D. J. S. (2020). A novel information processing in IoT based real time health care monitoring system. Journal of Electronics and Informatics, 2(3), 188-196.

[5] Belhadi, A., Kamble, S. S., Gunasekaran, A., Zkik, K., & Touriki, F. E. (2023). A Big Data Analytics-driven Lean Six Sigma framework for enhanced green performance: a case study of chemical company. Production Planning & Control, 34(9), 767-790.

[6] Sandhu, A. K. (2021). Big data with cloud computing: Discussions and challenges. Big Data Mining and Analytics, 5(1), 32-40.

[7] Mansour, R. F., Escorcia-Gutierrez, J., Gamarra, M., Díaz, V. G., Gupta, D., & Kumar, S. (2023). Artificial intelligence with big data analytics-based brain intracranial hemorrhage e-diagnosis using CT images. Neural Computing and Applications, 35(22), 16037-16049.

[8] Virnodkar, S. S., Pachghare, V. K., Patil, V. C., & Jha, S. K. (2020). Remote sensing and machine learning for crop water stress determination in various crops: a critical review. Precision Agriculture, 21(5), 1121-1155.

[9] Peng, D., Bruzzone, L., Zhang, Y., Guan, H., Ding, H., & Huang, X. (2020). SemiCDNet: A semisupervised convolutional neural network for change detection in high resolution remote-sensing images. IEEE Transactions on Geoscience and Remote Sensing, 59(7), 5891-5906.

[10] Khanra, S., Dhir, A., Islam, A. N., & Mäntymäki, M. (2020). Big data analytics in healthcare: a systematic literature review. Enterprise Information Systems, 14(7), 878-912.

[11] Bajaj, K., Sharma, B., & Singh, R. (2022). Implementation analysis of IoT-based offloading frameworks on cloud/edge computing for sensor generated big data. Complex & Intelligent Systems, 8(5), 3641-3658.

[12] Misra, N. N., Dixit, Y., Al-Mallahi, A., Bhullar, M. S., Upadhyay, R., & Martynenko, A. (2020). IoT, big data, and artificial intelligence in agriculture and food industry. IEEE Internet of things Journal, 9(9), 6305-6324.

[13] Li, Y., Zhang, Y., & Zhu, Z. (2020). Error-tolerant deep learning for remote sensing image scene classification. IEEE transactions on cybernetics, 51(4), 1756-1768.

[14] Lechner, A. M., Foody, G. M., & Boyd, D. S. (2020). Applications in remote sensing to forest ecology and management. One Earth, 2(5), 405-412.

Edited by: Zhigao Zheng
Special issue on: Graph Powered Big Aerospace Data Processing
Received: Nov 11, 2023
Accepted: Nov 24, 2023