Online Monitoring Method of Big Data Load Anomaly Based on Deep Learning

Cao-Fang Long and Heng Xiao(✉)

Sanya University, Sanya 572022, China
hh356398632@163.com, xiaoheng564@163.com

Abstract. In the process of monitoring the abnormal load of big data in network behavior, more network traffic resources are consumed, which leads to the low efficiency of its operation. Therefore, an on-line monitoring method for the abnormal load of big data in network behavior based on deep learning is proposed. The online monitoring model of load anomaly is established, the network data distribution is analyzed, and the adaptive random link configuration is adopted to improve the channel balance and the positioning ability of the abnormal load. The load anomaly is identified through the load pattern and the online monitoring is completed. The experimental results show that the proposed method consumes about 50% of the traffic of the traditional method, which can effectively reduce the traffic consumption and improve the utilization rate of network resources. This method is more suitable for online monitoring of big data load anomalies in network behavior.

Keywords: Deep learning · Network behavior big data · Online monitoring

1 Introduction

Deep learning is a new field in machine learning. Its motivation is to build and simulate the neural network of human brain for analysis and learning. It imitates the mechanism of human brain to interpret data, such as images, sounds and texts. The concept of deep learning comes from the research of artificial neural network. Multilayer perceptron with multiple hidden layers is a kind of deep learning structure. In-depth learning, by combining low-level features to form a more abstract high-level representation of attribute categories or features, to discover the distributed feature representation of data [1]. Machine learning is a subject that studies how computer simulate or realize human learning behavior to acquire new knowledge or skills and reorganize the existing knowledge structure to improve its own performance. In 1959, Samuel of the United States designed a chess playing program that has the ability to learn and can improve his own chess skills in continuous playing. Four years later, the program beat Samuel. In 1966, the program defeated an unbeaten American player for eight years. This program shows people the ability of machine learning, and puts forward many thoughtful social and philosophical problems. The research of machine learning is based on the understanding of human learning mechanism in physiology, cognitive science, etc., to establish the computational model or cognitive model of human learning process, to develop various learning theories and learning methods, to study

S. Liu and L. Xia (Eds.): ADHIP 2020, LNICST 348, pp. 452–462, 2021.
https://doi.org/10.1007/978-3-030-67874-6_42

the general learning algorithm and carry out theoretical analysis, to build a task-oriented learning system with specific application [2]. All kinds of human sensory organs are receiving a large number of data at any time. Some of these data come from the human body itself, some come from the external environment, but the brain can always obtain or extract the most important information that deserves attention. How to represent and analyze information efficiently and accurately is the core goal of machine learning. Through anatomical knowledge, experts in neuroscience have found the way in which human brain expresses information: unlike previous inferences, the cerebral cortex does not directly extract the eigenvalues of data, but uses a hierarchical network model constructed by the brain to analyze and filter the stimulus signals received by neurons, and then can obtain the characteristics and rules of the perceived data [3]. The above conclusions are analyzed by measuring the transmission time of sensory signals in retina, prefrontal cortex and motor nerve. In short, for the visual system, the human brain does not directly process the "first-hand data" obtained by the eyes, but recognizes objects according to the results of aggregation and decomposition. It can be seen that a clear hierarchy greatly reduces the amount of data that human visual system needs to process, and can retain the required information to the maximum extent. Deep learning is to simulate the human visual system and extract the essential characteristics of a large number of data with potentially complex structure rules.

With the rapid development of the Internet, the load of the access network increases gradually, and the data information generated by these loads becomes the data source of various analysis in the network. Although these data have become an important value data source for decision-making analysis. Theoretically, the more load is, the more valuable it is to obtain data samples. However, a large number of load access to the network, bringing new challenges to the stability and functionality of the entire network system [4]. Under the big data platform, the online load of the access network is limited by the total load, data bandwidth, computing capacity, response time, data carrying capacity of the accessible network of the platform. If the online load needs to exchange data with the platform, it must establish an effective connection with the platform. In order to prevent the waste of resources caused by the load occupying the data channel without effective communication, an effective method must be adopted to monitor the online load effectively and improve the utilization rate of resources in a timely and effective manner. The traditional online load monitoring method mainly relies on the communication between the server and the load as the standard to check whether the load is online. This method is also the main method to check whether the other party is online between the network points. However, this method has certain limitations, needs to occupy more network resources, which is not conducive to the rational application of data bandwidth. In this paper, the neural network in the field of deep learning is used to monitor the abnormal load of big data in network behavior.

2 Online Monitoring of Network Behavior Big Data Load Anomaly

2.1 Establishment of Online Monitoring Model for Abnormal Load

In order to realize the on-line detection of big data abnormal load in network behavior, first of all, data structure analysis and abnormal load information flow model construction are needed. Statistical characteristic analysis method is used to calculate the characteristic quantity of abnormal load data. According to the distribution attribute of the characteristic quantity, the detection model design of abnormal load data is realized, the output link model of optical fiber network is established, and the network is designed The channel equalization control model adopts the irregular triangle network model to build the big data sampling node distribution model of cloud computing optical fiber network, as shown in the Fig. 1:

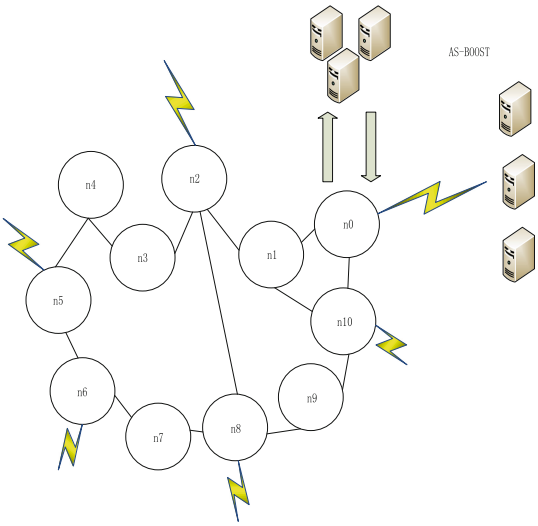

Fig. 1. Network data distribution

The network behavior big data load anomaly online detection model is composed of link layer, backbone node and sink node. The sampling data rule vector set of big data cluster head node in the network is distributed as follows:

$$\mathrm{SK_{i'}} = \begin{cases} \mathrm{K_i} & \text{if } i = 1 \\ New_{i'} & \text{otherwise} \end{cases} \tag{1}$$

Among them, $New_{i'} = (r_{i'1}, r_{i'2}, \ldots, r_{i'\mathrm{n}})$ represents the Source node set, uses the deep learning method to mine the big data abnormal load $\left\{M_h^{(0)}, h = 0, 1, \ldots, V-1\right\}$ in the network, sets the data flow of the time sequence sliding window, the sample clustering

weight is $\left\{a_h^{(0)}, h = 0, 1, \ldots, V-1\right\}$, replaces the association rule items on each sub window to the clustering result set, and obtains the initial value $\left\{S_h^{(0)}, h = 0, 1, \ldots, V-1\right\}$ of the data clustering center. In order to improve the accuracy of load monitoring, it is necessary to carry out channel equalization design. The iterative equation for constructing network channel link offset correction is as follows:

$$f_{ih}(v+1) = f_{ih}(v) + \mu_{MCMA} \frac{\alpha H_{MCMA}(v)}{\alpha f_{ih}(v)} \tag{2}$$

Among them, μ_{MCMA} represents the initial routing location of the network, $f_{ih}(v)$ represents the initial sampling value of the abnormal load information flow, and constructs the big data abnormal load information flow model.

The nonlinear time series analysis method is used to model the information flow of big data abnormal load in the network, and the output of offset load in the channel is obtained as follows:

$$a_i(v) = \sum_{h=1}^{W} g_{ih}(v)^T s_h(v) + n_i(v) \tag{3}$$

The channel model of big data transmission in the network is:

$$y_h(v) = \sum_{h=1}^{Q} f_{ih}(v)^T a_i(v) \tag{4}$$

Where, f_{ih} indicates the DNS load frequency of big data. In the current snapshot window, the number of data categories that satisfy the decision of data classification is accurate DB_{ih}. The tuples in the classification space of big data abnormal load in the network D[n + 1] are deleted. D[h] = D[h + 1], In the phase space supported by the limited data set, the vector quantization decomposition formula of big data abnormal load in the network can be described as follows:

$$\begin{aligned} \min_{\phi} \|Y - X\phi\| &= \min_{\phi} \left\| O^T Y - \sum N^T \phi \right\| \\ &= \min_{\phi} \left\| \begin{bmatrix} O_1^T \\ O_2^T \end{bmatrix} Y - \begin{bmatrix} \sum_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} N_1^T \\ N_2^T \end{bmatrix} \phi \right\| \\ &= \min_{\phi} \left\| \begin{bmatrix} O_1^T Y - \sum_1 N_1^T \phi \\ O_2^T Y \end{bmatrix} \right\| \\ &= \min_{\phi} \left\{ \left\| O_1^T Y - \sum {}_1 N_1^T \phi \right\| + X \right\} \end{aligned} \tag{5}$$

Among them, independent means the correlation coefficient X and ϕ aggregation coefficient of big data abnormal data in the network. The high-order statistics analysis

method is used to reconstruct the characteristics of abnormal load information flow. When the size set of big data abnormal load in the network tends to infinity, it can be discarded, that is:

$$\min_{\phi}\|Y - X\phi\| = \min_{\phi}\left\|O_1^T Y - \sum {}_1 N_1^T \phi\right\| \tag{6}$$

In the link model of big data transmission in the network, the spatial distribution cluster of the t load sampling node h in the dimension space i is obtained by the second iteration d(t) calculation, then:

$$\mathrm{d}_{ih}(\mathrm{t}) = |a_{ih}(\mathrm{t}) - j_{best}(\mathrm{t})| \tag{7}$$

Among them, the load time series is $a_{ih}(\mathrm{t})$ represented, and the fitness function is represented by j_{best}. The adaptive random link configuration method is used to detect the abnormal load and reorganize the data structure of the big data in the network, so as to improve the channel balance and the positioning ability of the abnormal load.

Then the exception detection and update strategy of the data i sampling node at the $(\mathrm{t}+1)$ moment is:

$$\begin{cases} n_{id}^{(\mathrm{t}+1)} = n_{id}^{t} + x_1 * u_1\left(q_{id}^{t} - a_{id}^{t}\right) \\ \quad + x_2 * u_2\left(q_{jd}^{t} - a_{id}^{t}\right) \\ a_{id}^{(\mathrm{t}+1)} = a_{id}^{t} + n_{id}^{(\mathrm{t}+1)} \end{cases} \tag{8}$$

Among them, $\{\mathrm{x}_1, \mathrm{x}_2\}$ the acceleration coefficient of single variable load detection is the random number between, which $\{\mathrm{u}_1, \mathrm{u}_2\}$ is $[0, 1]$ the lag detection coefficient of network big data abnormal load. m is the statistical characteristic quantity of abnormal load is extracted by time-frequency analysis method, and the correlation analysis is carried out according to the residual of detection model to build the network inspection regression analysis model. According to the residual of regression, the low inertia coefficient is adjusted adaptively

$$\begin{cases} d_{mean}(\mathrm{t}) = \dfrac{\left|\sum\limits_{h=1}^{v}\sum\limits_{i=1}^{d} d_{ih}(\mathrm{t})\right|}{v*d} \\ d_{mean}(\mathrm{t}) = |\max[d_{ih}(\mathrm{t})]| \\ k = \dfrac{|d_{\max}(\mathrm{t}) - d_{mean}(\mathrm{t})|}{d_{\max}(\mathrm{t})} \end{cases} \tag{9}$$

Among them, $d_{mean}(\mathrm{t})$ is the average particle distance, $d_{\max}(\mathrm{t})$ is the maximum particle moment and k is the clustering degree of the network big data distribution are used to detect the abnormal load of big data, and the value range $[0, 1]$ is the statistical characteristic of the abnormal load extracted from them, and the abnormal data is diagnosed according to the feature extraction results.

2.2 Abnormal Data Diagnosis

In the phase of abnormal data diagnosis, the monitoring program will analyze the time sequence pattern of the obtained network data based on a small number of network characteristic parameters selected in advance, match the fault characteristics in the network database, and identify the suspicious network parameter characteristics based on the matching degree [5]. Compared with the traditional network fault monitoring stage, the method proposed in this chapter improves the time sequence of the original method, and more effectively predicts the precursor of the fault from the perspective of time. In the second stage, if the matching degree in the first stage is large, the model diagnosis stage procedure will be triggered [6]. The fault diagnosis model trained by the program using historical data will reapply the detailed parameter information of the network to the operation and maintenance management system. OAM will input the detailed parameter information of the network in the near time period into the model through data processing, and output the classification result (Fig. 2):

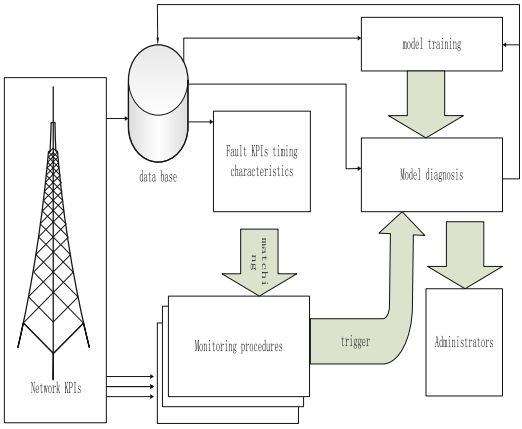

Fig. 2. Two stage fault diagnosis block diagram

In the traditional wireless cellular network, users at the boundary of base station are easily interfered by the neighbor base station. This effect can be based on frequency multiplexing technology or power control technology to reduce the degree of mutual interference [7]. But in the layered heterogeneous network, the interference generally includes the same layer network interference and the cross layer network interference. Cross layer network interference occurs between high-power macro base station and low-power base station. The low-power base station has the characteristics of intensive deployment. In a macro cell, there may be hundreds of home cell deployments. Moreover, due to the weak planning of the deployment, too many low-power base station deployments may make Acer station users included in the coverage of low-power base stations. On the one hand, the uplink signal of the user in the Acer Service will affect the performance of the low-power base station; on the other hand, the downlink signal of the low-power base station will also interfere with the user

experience in the Acer station. The interference in the same layer is mainly reflected in the low-power base stations. The spatial distribution characteristics of low-power base stations are diverse, resulting in more complex interference environment [8]. Due to the weak planning of deployment, there will be overlapping coverage, which makes the interference everywhere. The complex interference environment is the main factor to reduce the performance of wireless network system. It not only reduces the network throughput, limits the network spectrum utilization, but also affects the stability of wireless link, causing frequent drop of users. Therefore, although the user access information can reflect the base station load from one point of view, it does not show the information of base station failover. Moreover, the load of low-power base stations is relatively small, and the user access of neighbor base stations may cause the load of neighbor base stations to be too heavy, again affecting the performance of neighbor base stations [9, 10]. To solve the problem of all users switching caused by the faulty base station, more comprehensive base station information is needed to represent, which can quickly find the fault of the original base station before the secondary performance pollution of the base station, so as to realize the online monitoring of the network behavior big data load anomaly.

2.3 Online Monitoring

During the operation of the network, various load patterns are obtained by monitoring the load vector sequence, each of which corresponds to a measurement space, as shown in the Fig. 3:

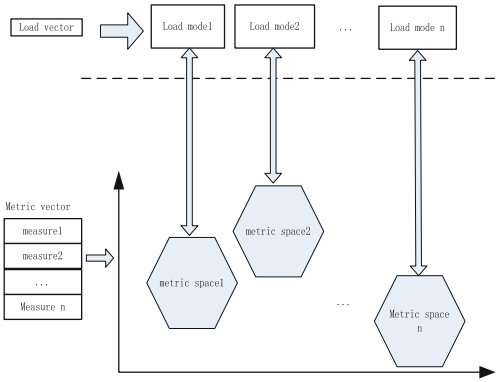

Fig. 3. Relationship between load pattern and metric space

As shown in the figure above, the corresponding relationship between the load pattern and the measurement space is represented by the stable operation state, which is defined as follows:

$$SS_i = \{KM_i, D_i, WS_i\} \tag{10}$$

Among them, KM_i is the center of the cluster i; D_i is the covariance matrix of all dimensions of the load vector in the cluster i; WS_i is the measurement space of the load pattern, which is composed of the set of measurement vectors $\{WN_1, WN_2, \ldots, WN_i, \ldots\}$ and the distance Euclidean.

In the process of running, we match the load vector of the running state with the known load pattern to obtain its corresponding measurement space, and use the local exception factor to calculate the exception score of the current measurement vector according to the measurement space [11, 12]; when determining the exception, we use the test to inspect each measurement separately to locate the exception measurement, and the exception monitoring method is shown in the Fig. 4:

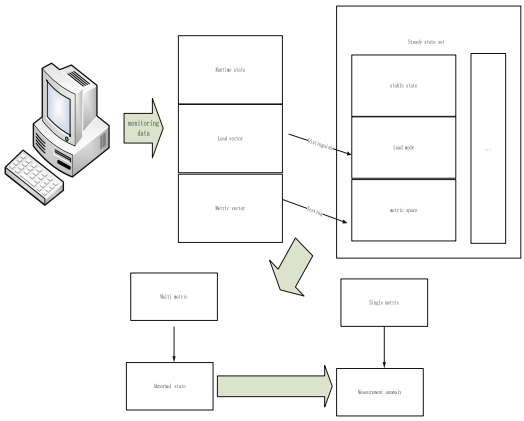

Fig. 4. Abnormal monitoring

By monitoring the changes of measurement, we can detect the abnormal state in the process of network behavior big data application running. At the same time, we can identify the abnormal load caused by the severe load fluctuation through the load pattern, and complete the online monitoring. In order to test the application effect of the proposed method, the following experiments are designed.

3 Case Simulation

3.1 Experiment Preparation

In view of the special problems that need to be solved in the process of online monitoring of big data load abnormality of network behavior, there are also special requirements for the construction of experimental environment. The construction of software environment for the experiment is shown in the table below (Table 1):

Table 1. List of software required for simulation experiment operation

	Operating system	Data base	Application software
Data server	Windows Server 2016	Oracle 11 g	ArcSDE 10.3
GIS server	Windows Server 2016		ArcGIS Server10.2
Web server	Windows Server 2016		JDK 1.6
Desktop client	Windows 10		IE browser 11.0
Mobile client	Android 10.0	SQLite	

At this point, the experimental preparation is completed, and the experimental platform can meet the needs of experimental operation. The specific results and analysis are as follows.

3.2 Result Analysis

In order to verify the performance of this method in load monitoring, cloudsim simulation platform is used to simulate the algorithm performance, and the data sample is 10000 loads. Firstly, 10000 sample sets are set up, all of them are connected to the data platform, then the load condition is adjusted continuously within 60 min, and some of the load is forced to lose the connection manually, so as to test the performance of this method in abnormal load monitoring. In 60 min, adjust the number of dropout load continuously. The upper limit of dropout load is 100, accounting for 1% of the total load. Test the adaptability of the algorithm in this paper. The simulation results are shown in the Fig. 5.

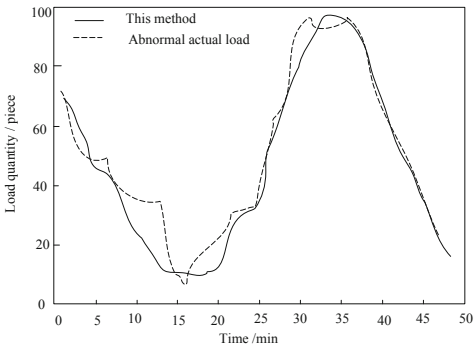

Fig. 5. On line monitoring of abnormal load in this method

As can be seen from the figure, this method can detect the abnormal load of the platform well. In the first 20 min, the two lines are coincident and fit well; but in the period of 20–30 min, some of the load is not detected; in the last 30 min, the method in this paper is better to detect the load.

In order to verify the advantages of the algorithm in data flow, compare the traditional method with the method in this paper, respectively simulate the monitoring flow consumed by the two monitoring algorithms when the number of loads is 5000, 10000, 15000, 20000 and 25000, and the results are shown in the figure below.

As you can see from Fig. 6, the monitoring flow will increase as the number of monitoring loads increases. When the load is 5000, the monitoring process of traditional method 1 is about 370 MB, and that of this method is 180 MB. The traditional method consumes about twice as much traffic as this method. Therefore, this method has an absolute advantage in the use of network affine resources, that is, the operation efficiency is high.

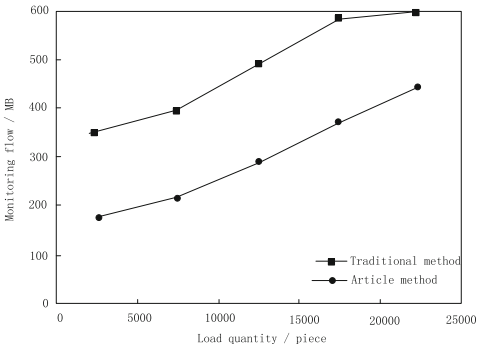

Fig. 6. Comparison of consumed flow

4 Summary and Prospect

When using the big data platform, workers should be clear: the online load of the access network is easily interfered by many factors, such as data bandwidth and computing power, etc., which requires workers to do data exchange processing between the online load and the platform to ensure that it can keep close contact with the platform. In order to avoid the problem of occupying data channels and wasting resources excessively, the staff should further monitor the online load in a practical way and make full use of resources.

Under the background of the rapid development of the Internet, the amount of network access load is increasing. The main analysis data of the network is generated from the resources of the load. And on this basis, referring to the specific changes of the load quantity in the adjacent time period, set up a load forecasting model that conforms to the status quo, and make a judgment on the load status according to the load status quo. For a period of time when the load changes greatly, the staff should make clear the load state through deep learning, and make forward and reverse operations on it. Although this kind of data is an essential resource for decision-making and analysis, and the more load, the higher data practicability, but a large number of loads connected with the network, the stability of the network system is undoubtedly a challenge.

5 Concluding Remarks

This paper proposes an online monitoring method of big data load anomaly based on deep learning. The experimental results show that the method in this paper can detect the load better within the limit of 30 min, and the load is 180 MB when the load is 5000, which shows that the method has absolute advantages in operation efficiency, which can better monitor the real connection of big data platform load and effectively reduce the occupation With the off-line load of network bandwidth, the efficiency of big data platform is improved, and less traffic is consumed in the use of network resources. Therefore, deep learning has strong applicability, online monitoring the application of big data load anomaly in network behavior.

References

1. Wang, Y., Tang, J.: Deep learning-based personalized paper recommender system. J. Chin. Inf. Process. **32**(04), 114–119 (2018)
2. Tang, C., Ling, Y., Zheng, K., et al.: Object detection method of multi-view SSD based on deep learning. Infrared Laser Eng. **47**(01), 302–310 (2018)
3. Zhang, Y., Li, M., Han, S.: Automatic identification and classification in lithology based on deep learning in rock images. Acta Petrol. Sin. **34**(02), 333–342 (2018)
4. Li, X.: Research on big data online load abnormal monitoring technology based on wavelet neural network. Mod. Electron. Tech. **42**(11), 95–97 (2019)
5. Liang, L., Li, J.: On-line load abnormality monitoring technology for large data based on wavelet neural network. Adhesion **40**(09), 94–96 + 116 (2019)
6. Chuili, H.U.: Distributed internet resources load balancing distribution simulation. Comput. Simul. **35**(07), 241–244 (2018)
7. Liu, S., Fu, W., He, L., et al.: Distribution of primary additional errors in fractal encoding method. Multimed. Tools Appl. **76**(4), 5787–5802 (2017). https://doi.org/10.1007/s11042-014-2408-1
8. Lu, Y., Zhang, T., He, E.: Probabilistic routing-based data fusion method in multi-source and multi-sink WSNs. Trans. Microsyst. Technol. **38**(07), 53–56 (2019)
9. Liu, S., Liu, G., Zhou, H.: A robust parallel object tracking method for illumination variations. Mob. Netw. Appl. **24**(1), 5–17 (2019). https://doi.org/10.1007/s11036-018-1134-8
10. Sun, L., Yu, K.: Research on big data analysis model of library user behavior based on Internet of Things. Comput. Eng. Softw. **40**(06), 113–118 (2019)
11. Liu, S., Liu, D., Srivastava, G., et al.: Overview and methods of correlation filter algorithms in object tracking. Complex Intell. Syst. (2020). https://doi.org/10.1007/s40747-020-00161-4
12. Lu, M., Liu, S.: Nucleosome positioning based on generalized relative entropy. Soft. Comput. **23**, 9175–9188 (2019). https://doi.org/10.1007/s00500-018-3602-2