Research on Host Intrusion Detection Method Based on Big Data Technology

Lei Ma(✉) and Hong-xue Yang

Beijing Polytechnic, Beijing 100016, China
malei235@tom.com

Abstract. When the host runs a large number of applications at the same time under normal activities, the abnormal probability value of the host after the fusion of evidence is large, resulting in false alarms, resulting in a reduction in the final detection accuracy of the detection method. A host intrusion detection method based on big data technology. Using big data processing intrusion detection index weight, sliding window is introduced. According to the number of times of host resource availability anomaly in the time window, the value of anomaly probability is controlled, the index anomaly closed value is determined, and the availability anomaly threshold is set to realize host intrusion detection. The experiment builds a data collection platform and compares the two traditional detection methods with the detection methods studied in the paper. The results show that the detection accuracy of the proposed detection method is about 98%, and the detection of host intrusion behavior is more accurate and the detection time is shortened.

Keywords: Big data technology · Intrusion behavior · Outlier probability · Detection accuracy

1 Introduction

With the rapid development of Internet technology, a lot of data information is produced in all aspects of production and life, which makes the data sources more and more extensive. At the same time, network security management is also facing severe challenges. There are more and more hacker attack channels, more Trojan horses and virus technologies. The speed of network security analysis data increases exponentially and the lag of data analysis speed increases A network security vulnerability [1]. Big data technology has the ability to quickly obtain valuable information from a large number of data with complex structure and various types. It can reveal the content and change trend that can't be seen by traditional means. It is a hot topic in the current academic, industrial and even national governments. Big data technology brings new opportunities and challenges to the development of the information security industry. In order to better play the role of big data and reduce the damage of data caused by cyber attacks, the research of intrusion detection technology based on big data is urgent, so that big data The safe and stable interaction of information data in this era.

In various industries and related fields, relying on the background of the network era, big data provides a good platform and effective channels for resource sharing and

S. Liu and L. Xia (Eds.): ADHIP 2020, LNICST 347, pp. 438–449, 2021.
https://doi.org/10.1007/978-3-030-67871-5_39

data exchange. However, with the increase of data volume and centralized storage of data, the security protection of massive data becomes more difficult, and a large number of centralized storage and processing of data inevitably increases the number of users According to the risk of leakage. The targets of hackers are more related to data clusters such as financial institutions, large companies, campus networks, etc. Once successfully attacked, hackers will obtain huge wealth from them. This year's ransomware attacked finance, government, and enterprises [3].

With the advent of the era of big data, hackers' attack methods are also changing, showing the development trend of covert attack methods. Analysis of its source is mainly due to the sudden increase in the amount of data and the increasingly close relationship between the data, which not only makes it difficult to detect hacker attacks, but also brings a wider range of harm through the connection between data and information. Traditional intrusion detection technology collects some network status information by placing multiple detectors in the network, and then the central control center analyzes and processes the information. However, in the face of large-scale, heterogeneous network environment and distributed cooperative attack, it is not enough. The main reasons are: first, the workload of the management control center is too large, and there are problems with the operation of the system; second, there is a certain delay in network transmission, and the transmitted information data cannot be transmitted to the management control center in time; third, there are platform differences in heterogeneous networks, making the analysis system face many difficulties. In view of these reasons, the distributed intrusion detection system came into being, and once became a hot spot in the field of intrusion detection [3]. Therefore, this paper proposes a host intrusion detection method based on big data technology.

2 Research on Host Intrusion Behavior Detection Method Based on Big Data Technology

2.1 Index Weight of Intrusion Detection in Big Data Processing

Big data technology intrusion behavior detection index weights adopt the objective weight assignment method to form the actual data of each host resource characteristic index in the decision plan. It has the advantages of objectivity and strong mathematical theoretical basis. It is determined by the analytic hierarchy process After the subjective weights of various indicators of the host's resources, the entropy weight method is used to determine its objective weights. When the information entropy takes the maximum value, the probability of the corresponding set of states appearing has an absolute advantage $H(X)$. The description formula of $H(X)$ is:

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i (0 \le p_i \le 1, \sum_{i=1}^{n} p_i = 1) \tag{1}$$

In the above formula: p_i represents the extreme value of entropy and i represents the quantity. The information entropy solves the problem of measuring the amount of information and is a measure of the uncertainty of the system state. That is, the amount

of information required to understand the uncertainty can be used to eliminate uncertainty. How much to express, entropy is a commonly used indicator to measure the regularity of disordered data. Entropy weight is the relative intensity of each index in the sense of competition when all kinds of detection indexes are determined after given factor set and measurement set [4]. If the degree of variation of the index value is smaller, the corresponding information entropy value is larger, the amount of information provided by the index is smaller, and the weight of the index is also smaller. Therefore, information entropy is an objective method to allocate the weight, which is determined by calculating the entropy value. In essence, it is to select the best factor to reflect the availability of host resources. The application of entropy weight method can eliminate the artificial interference in the calculation of each index weight in AHP, and make the evaluation result more real.

For the abnormality of the i index R, denoted by r, and then give a quantitative definition of the abnormality, divided into 4 levels, as shown in the following Table 1:

Table 1. Quantification level of abnormality degree

Abnormal degree	Abnormality	Quantized value
r_1	The indicator usage increment exceeds 0–15%	0.1
r_2	The use increment of this indicator is more than 16–50%	0.25
r_3	The use increment of this indicator is more than 51–70%	0.7
r_4	The use increment of this indicator is more than 71–100%	1

According to the grade index shown in the table above, and according to the calculation formula of entropy, after normalizing the relative importance of index R_i, the calculation is expressed as:

$$e_i = -\frac{1}{\ln n}\sum_{i=1}^{n} r_i \log r_i \tag{2}$$

In the above formula, when the values of $r_i(i = 1, 2, \ldots, n)$ are equal, the entropy value e_i is at most 1, so when $0 \le e_i \le 1$, $\ln n$ is the maximum entropy value when all are equal. When the entropy value is maximum, the contribution of this indicator to the detection result is the smallest, so the weight of the detection factor R_i can be determined using the $1 - e_i$ metric process. Therefore, it is normalized to obtain the objective weight of the detection factor R_i:

$$w_i = \frac{1 - e_i}{n - \sum_{i=1}^{n} e_i}, 0 \le w_i \le 1, \sum_{i=1}^{n} w_i = 1 \tag{3}$$

In the above formula, w_i represents objective weight. The weight obtained by AHP reflects the subjective preference of decision-makers, and the weight calculated by entropy weight method reflects the objective relationship between each host resource index. In order to give consideration to the expert's experience judgment of attributes, at the same time, it strives to reduce the subjective randomness of subjective weight. Use the linear weighting method to determine the comprehensive weight of the attribute, namely:

$$w_i = \mu w_{si} + (1 - \mu) w_{oi}, (i = 1, 2, \ldots n) \tag{4}$$

In the formula, w_i is the comprehensive weight, w_{si} is the subjective weight of the index determined by the analytic hierarchy process, and w_{oi} is the objective weight of the index determined by the entropy weight method. μ is the subjective preference coefficient, $1 - \mu$ is the objective preference coefficient, and $0 \leq \mu \leq 1$. When μ is less than 0.5, the proportion of subjective weight in the overall weight is relatively large, and vice versa. In general, the specific value of μ is given by the decision maker according to preferences, and it can be proved that μ can obtain the optimal value and optimize the change results, as shown in the following Fig. 1:

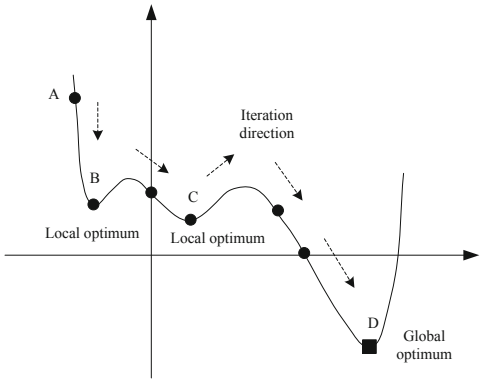

Fig. 1. Optimize change

According to the change process shown in the figure above, suppose S represents the host information system, w_{si} and w_{oi} are the subjective weight and objective weight of indicator R_i respectively, and w_i represents the combined weight of the two. An optimization model is established for the above indicators, and the calculation formula is as follows:

$$\min\left\{\sum_{i=1}^{n}\left[\mu\left(\frac{1}{2}(w_i - w_{st})^2\right) + (1 - \mu)\left(\frac{1}{2}(w_i - w_{oi})^2\right)\right]\right\} \tag{5}$$

Using the above formula to optimize the above weight value calculation process, according to the intrusion behavior detection index weights, determine the abnormal closed value of each index to realize the detection of host intrusion behavior.

2.2 Determination of Abnormal Closed Value of Each Index

When determining the abnormal closed value of each index, first of all, you need to obtain the host resource availability profile under normal circumstances, and use this to determine the abnormal value of the host resource index [5]. We collect data samples on several hosts in the laboratory. Host resource availability indicators are collected once a minute, 1440 times in a continuous 24 h. In order to ensure the objectivity of experimental data, the normal operation of the host computer is ensured in the process of data acquisition. For each index data collected, we first calculate the average m and standard variance c, and then calculate the corresponding normal value range H average and standard variance calculation formula of the index:

$$\begin{cases} m_i = \sum_{i=1}^{n} \frac{r_i}{n} \\ c_i = \sqrt{\sum_{i=1}^{n} \frac{(r_i - m_i)^2}{n-1}}, i = (1, 2, \ldots, 12) \end{cases} \tag{6}$$

In the above formula, the normal value range can be calculated as follows:

$$H_i = m_i - d \times c_i, m_i + d \times c_i \tag{7}$$

In the above formula, A represents d constant. It can be determined according to the actual situation of different resource indicators. For the threshold value of target host resource availability judgment and the setting of the closed value of host security status exception, due to the variety of exception status, it is impossible to accurately determine the corresponding value [6].

In actual detection, some attacks are manually launched on the target host, and then the corresponding threshold is calculated according to the sampled data. Of course, the corresponding threshold can also be set based on the experience of the administrator to determine the pairwise comparison matrix, where the target layer judgment matrix is A, and the matrix can be expressed as:

$$A = \begin{bmatrix} 1 & \frac{1}{2} & 4 & 3 & 3 \\ 2 & 1 & 7 & 5 & 5 \\ \frac{1}{4} & \frac{1}{7} & 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{5} & 2 & 1 & 1 \\ \frac{1}{3} & \frac{1}{5} & 3 & 1 & 1 \end{bmatrix} \tag{8}$$

In the indicator layer, there is only one indicator of CPU utilization and the indicator matrix of network resources is B_3, so it directly inherits the result of the judgment matrix of the target layer. The calculation matrix can be expressed as:

$$B_3 = \begin{bmatrix} 1 & 3 \\ \frac{1}{3} & 1 \end{bmatrix} \tag{9}$$

Based on the above formula, the storage resource index is B_2, and the process/thread index matrix is B_5. The values of the two indexes are equal. The calculation formula is as follows:

$$B_2 = B_5 = \begin{bmatrix} 1 & 3 & 5 \\ \frac{1}{3} & 1 & \frac{1}{5} \\ \frac{1}{5} & \frac{1}{3} & 1 \end{bmatrix} \tag{10}$$

The resource matrix can be recorded as B_4:

$$B_4 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \tag{11}$$

Use the comparison matrix given above based on expert experience to calculate the weight vectors of the hierarchical single ranking, and after the consistency test, use big data technology to build a recognition framework, that is, filter out security events per unit time and make corresponding statistics and records, Construct the evidence collection:

$$\bigcup E_i (i = 1, 2, \ldots, n) \tag{12}$$

According to the above formula, we use the evidence BPA method to allocate the credibility of the data in the host, calculate the trust function and likelihood function of each data, and use the Dempster synthesis method to calculate the joint credibility evaluation function, trust function and likelihood function of the known evidence. If there is a new intrusion in the detection cycle, continue the data fusion process until the end of the detection cycle, build a distribution function according to the comprehensive trust level, calculate the trust level of the intrusion host data, and use the small value of the trust level value as the index abnormal closed value. Use these abnormal closed values to realize the detection of host intrusion behavior [7].

2.3 Implement Host Intrusion detection

There are many parameters that can be used for the statistical characteristics of host resource availability. Drawing on previous work, and through a large number of experimental analysis, five primary indicators are selected to form a host resource availability measurement system. Each primary indicator also includes a corresponding secondary Indicators [8, 9]. Because the change of each indicator can better reflect the dynamic characteristics of the application program's demand for the availability of host resources, and the incremental information can also reduce the amount of calculation and memory utilization, all the secondary indicators in the host resource indicators

selected by this method, except the CPU utilization, represent the increment in unit time. "Increment" refers to the absolute value of the change of the indicator relative to the last unit time. The five first-level indicators are the total statistics of storage resources, computing resources, bandwidth resources, processes/threads, and IO resources. Corresponding to its index level and meaning in the host, as shown in the following Table 2:

Table 2. Host availability index and meaning

First level indicators	Secondary index	Variable representation	Specific meaning
Computing resources	CPU	CPUUse	CPU Usage rate
Storage resources	physical memory	phyMem	Physical memory usage increment
	Virtual Memory	virtMem	Virtual memory usage increment
	Memory usage	memUse	Approved memory usage increment
Internet resources	Send traffic	bandSend	Send packet increment
	Receive traffic	bandReci	Receive data packet increment
Process/thread count	Process	Process	Process number increment
	Thread	Thread	Thread increment
	Handle	Count	Handle number increment
IO resources	IO reading	IO read	IO read increment
	IO write	IO write	IO write increment
	IO others	IO other	IO other increments

According to the above weight determination process, each index shown in the above table is assigned to the first and second level indexes, and the mathematical model of resource availability evaluation of the target host is established, with the calculated abnormal closed value as the evaluation index, Set the threshold of the total intrusion behavior data of the switchboard to L. When the abnormal closing value is greater than or equal to L, there is an abnormal intrusion and a sliding window is introduced, According to the number L of occurrences of host resource availability abnormalities in the time window, it is determined whether the current host security status is abnormal, that is, the abnormal status of the host at the current time is determined according to the number of host resource availability abnormalities within the latest window. among them:

$$p = \sum_{t-\Delta t}^{\Delta t} c \tag{13}$$

Among them, c is the number of host resource intrusions per unit time, Δ is the length of the time window, t is the intrusion time, $A(s)$ is used to indicate the current state of the host, 1 is abnormal, and 0 is normal, then the security status frequency is closed φ:

$$A(s) = \begin{cases} 1, p \geq \varphi \\ 0, p < \varphi \end{cases} \tag{14}$$

In the above formula, when $p < \varphi$, it means that the host is safe and there is no intrusion behavior; when $p \geq \varphi$, there is intrusion behavior. Based on the threshold value of host resource availability exception, judge the host resource availability exception within the window time, and judge the host security status exception at the current time according to the closed value of the host security status exception frequency [10, 11]. The thresholds L and φ are determined by the network administrator according to the experience and the requirements of the host security state. The research on the host intrusion detection method based on big data technology is finally completed.

3 Simulation Experiment

3.1 Experiment Preparation

Three computers with the same parameters are prepared for the experiment. The parameters of the computers are as follows (Table 3):

Table 3. Experimental computer parameters

No. Name parameter	No. Name parameter	No. Name parameter
1	Memory	8 GB DDR4 RAM
2	Storage	Solid state drive SSD
3	Processor	Amd Ruilong 5 3580u mobile processor integration
4	Graphics card	Amd radeon TM Vega graphics card
5	interface	Usb-c, usb-a, surface connect port

According to the parameters shown in the table above, build a data collection platform, as shown in the following Fig. 2:

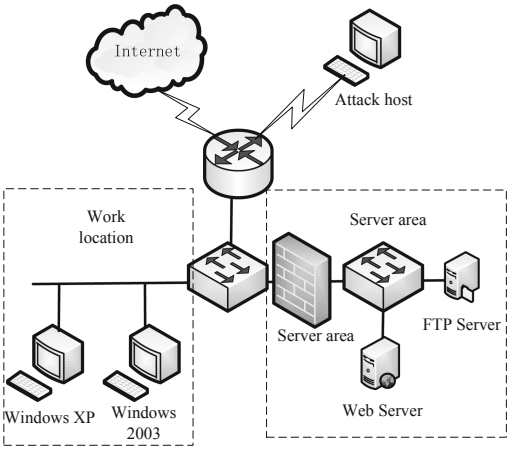

Fig. 2. Data acquisition platform

According to the platform structure shown in the figure above, simulate host intrusion behavior, deploy intrusion detection points, collect all host data and intrusion data, as shown in the table below (Table 4):

Table 4. All data and intrusion data in the host

Number of layers	32-64-256-128-64-32	32-1152-2
1	8.0365	0.8286
2	4.8646	0.7035
3	3.0567	0.6154
4	6.8237	0.5525
5	3.0342	0.5080
6	4.5420	0.4864
7	7.3056	0.4797
8	7.1946	1.3818
9	6.1387	0.9705
10	5.1188	0.5335

In the table above, 32-64-256-128-64-32 represents the data in all levels of the host, and 32-1152-2 represents the external intrusion data of the host. Network intrusion detection method based on improved majorcluster clustering (traditional behavior detection method 1) and host intrusion detection method based on data mining (traditional behavior detection method 2) are used respectively, and the proposed big data technology-based intrusion detection method proposed in this paper Host intrusion detection methods are tested to compare the performance of the three methods [12, 13].

3.2 Analysis of Results

Based on the above processing, the host intrusion detection time is obtained by iterating the data shown in the table above, as shown in the following figure:

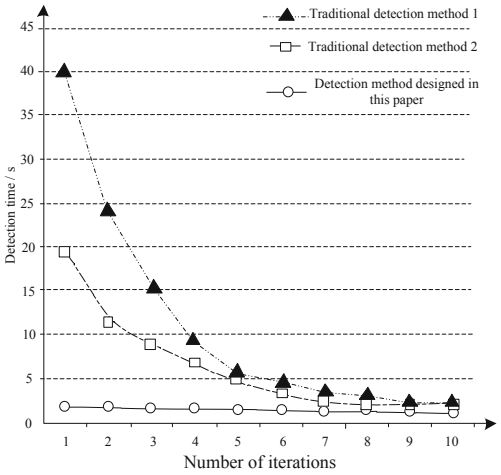

Fig. 3. Comparison of test time results

According to Fig. 3, with the increasing number of iterations, the detection time of the three detection methods gradually decreases. The host intrusion detection time of the traditional behavior detection method 1 is reduced from 40 s to 3S, and the host intrusion detection time of the traditional behavioral behavior detection method 2 is reduced from 15 s to 3S. However, the host intrusion detection time of the proposed method is relatively stable, which is reduced from 3S to 2S The time was the lowest among the three methods.

In this experimental environment, the detection accuracy of the three detection methods is calculated, and the experimental results are as follows Fig. 4:

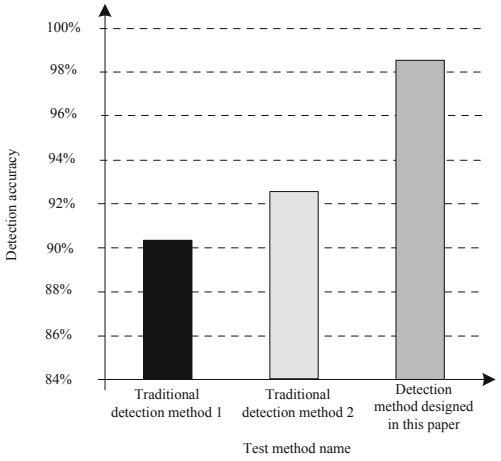

Fig. 4. Experimental results of accuracy of three detection methods

From the accuracy calculation results shown in the figure above, we can see that the three host intrusion behavior detection methods all have high test accuracy, but there are still certain differences in the size of the values. The accuracy of traditional detection method 1 is about 90%, the accuracy of traditional detection method 2 is about 92%. The detection accuracy of the test method studied in this paper is about 98%. This method has high sensitivity to host intrusion data to ensure the accuracy of intrusion data detection and is more suitable for the actual host intrusion detection.

4 Conclusion

In accordance with the development of big data technology, with the continuous expansion of data scale, the continuous improvement of data throughput and the requirements of multi hosts and multi network segments, the intrusion detection technology based on big data will play a huge role in the specific application platform in this field. The distributed intrusion detection method proposed in this paper improves the shortcomings of traditional intrusion detection system. It not only makes DIDS technology develop rapidly, but also provides safe and reliable protection for large-scale network.

References

1. Lakshmanaprabu, S.K., Shankar, K., Rani, S.S., et al.: An effect of big data technology with ant colony optimization based routing in vehicular ad hoc networks: towards smart cities. J. Cleaner Prod. **217**, 584–593 (2019)
2. Chen, M.-Y., Lughofer, E.D., Polikar, R.: Big data and situation-aware technology for smarter healthcare. J. Med. Biol. Eng. **38**(6), 845–846 (2018)
3. Ge, M., Bangui, H., Buhnova, B.: Big data for internet of things: a survey. Future Gener. Comput. Syst. **87**, 601–614 (2018)
4. Turner, C., Gill, I.: Developing a data management platform for the ocean science community. Mar. Technol. Soc. J. **52**(3), 28–32 (2018)
5. Rycarev, I.A., Kirsh, D.V., Kupriyanov, A.V.: Clustering of media content from social networks using bigdata technology. Comput. Opt. **42**(5), 921–927 (2018)
6. Watson, H.J.: Update tutorial: big data analytics: concepts, technology, and applications. Commun. Assoc. Inf. Syst. **44**(1), 364–379 (2019)
7. Zeng, W., Xu, H., Li, H., et al.: Research on methodology of correlation analysis of sci-tech literature based on deep learning technology in the big data. J. Database Manage. **29**(3), 67–88 (2018)
8. Shuai, L., Weiling, B., Nianyin, Z., et al.: A fast fractal based compression for MRI images. IEEE Access **7**, 62412–62420 (2019)
9. Xue, Y., Feng, H.: Path analysis of forest carbon sequestration on poverty alleviation papermaking company innovation based on big data analysis. Paper Asia **35**(1), 28–32 (2019)

10. Fu, W., Liu, S., Srivastava, G.: Optimization of big data scheduling in social networks. Entropy **21**(9), 902 (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. Huda, M., Maseleno, A., Atmotiyoso, P., et al.: Big data emerging technology: insights into innovative environment for online learning resources. Int. J. Emerg. Technol. Learn. **13**(1), 23–36 (2018)
13. Lu, M., Liu, S.: Nucleosome positioning based on generalized relative entropy. Soft Comput. **23**(19), 9175–9188 (2018). https://doi.org/10.1007/s00500-018-3602-2