Research Article

Yan Zhang*, Arshpreet Kaur, Vishal Jagota, and Rahul Neware

Study on data mining method of network security situation perception based on cloud computing

https://doi.org/10.1515/jisys-2021-0264
received December 02, 2021; accepted March 30, 2022

Abstract: In recent years, the network has become more complex, and the attacker's ability to attack is gradually increasing. How to properly understand the network security situation and improve network security has become a very important issue. In order to study the method of extracting information about the security situation of the network based on cloud computing, we recommend the technology of knowledge of the network security situation based on the data extraction technology. It converts each received cyber security event into a standard format that can be defined as multiple brochures, creating a general framework for the cyber security situation. According to the large nature of network security situation data, the Hadoop platform is used to extract aggregation rules, and perform model extraction, pattern analysis, and learning on a network security event dataset to complete network security situation rule mining, and establish a framework for assessing the state of network security. According to the results of the federal rule extraction, the level of network node security risk is obtained in combination with signal reliability, signal severity, resource impact, node protection level, and signal recovery factor. A simulation test is performed to obtain the intrusion index according to the source address of the network security alarm. Through the relevant experiments and analysis of the results, the attack characteristics obtained in this study were obtained after manually reducing the network security event in the 295 h window. The results show that after the security event is canceled, the corresponding window attack index decreases to 0, indicating that this method can effectively implement a network security situation awareness. The proposed technique allows you to accurately sense changes in network security conditions.

Keywords: data mining technology, network, security situation, cloud computing, security situational awareness

1 Introduction

In recent years, networks have become more complex, and the ability of attackers to attack is gradually increasing. How to properly understand the network security situation and improve network security has become a very important issue. Situational knowledge is widely used in many important fields, such as aviation, emergency services, and the military. This study uses it in the field of network security and suggests a technology for knowledge of network security situations based on data mining [1]. Understanding the network security situation is the foundation of network management and improves network

* **Corresponding author: Yan Zhang,** Department of Information Engineering, ShiJiaZhuang University of Applied Technology, Shijiazhuang, Hebei, 050081, China, e-mail: zhangyan9913@126.com
Arshpreet Kaur: GNA University, Village Hargobindgarh, Phagwara, Punjab, India, e-mail: arshpreetofficial@gmail.com
Vishal Jagota: Department of Mechanical Engineering, Madanapalle Institute of Technology & Science, Madanapalle, AP, India, e-mail: vishaljagota@mits.ac.in
Rahul Neware: Department of Computing, Mathematics and Physics, Høgskulen på Vestlandet, Bergen, Norway, e-mail: rane@hvl.no

security early warning capabilities by collecting and processing network security changes in a timely manner [2]. An attack on the network causes a flood of early warning messages that can be ignored. There are many types and it is difficult to determine the relationship between them. How to correctly define offensive behavior is an issue that needs to be addressed in a safe environment. To address these issues, data extraction-based network security technology is proposed [3–5].

The concept of the situation arose through the study of the human factor in spaceflight and was intensively studied in areas such as air traffic control, nuclear reaction control, and military battlefields. As network security has become a key issue in the network security industry, understanding network security has become an important tool for ensuring the security of the network environment and suppressing network security threats hidden in communications. Cyber situation knowledge is the ability to detect security risks through a comprehensive analysis of the dynamics of the external environment. Knowledge of network conditions is an effective way to identify, analyze, and respond to potential threats to network security, and is a big data as a core technology. Data mining is the process of automatically extracting dangerous data that may be hidden in large amounts of network data. The extraction of network potential risk information analyzes the commonalities and differences between network potential risk data and, based on this, finds possible rules between network potential risk data. By preparing network data, finding abnormal network data, and summarizing abnormal network data rules, it is possible to determine whether abnormal network data may be dangerous.

1.1 Framework of network security situation perception technology architecture

Obtaining information about the situation in the field of network security refers to the process of obtaining valuable and important information about the state of network security from large-scale data sources, and it is the basis for quantitative perception and prediction of the situation. The obtained objects and results directly affect the form and accuracy of quantitative perception. At present, the research in this technical field is still in its infancy, and there are few corresponding research literature, but the work of feature recognition, classification analysis, and clustering has been carried out in the early stage. Network security situational awareness based on data mining can sense the specific changes in network security situation through the iteration of network potential danger data, so as to improve the accuracy of network security situational awareness.

The network situation refers to the network operation status and change trend composed of factors such as the working conditions of different network devices, network behaviors, and user behaviors [6]. Network situation awareness is the collection of security factors that can lead to changes in the network situation in a complex network environment, and prediction of the future situation of the network [7]. Use big data technology to integrate different types of perception data sources, and use data mining technology to convert different network security data that are disorderly and seemingly unrelated into intuitive information to complete network security situation perception. Awareness of the network security situation mainly includes the following three important stages. The first stage is to collect different types of security data such as terminals, borders, and applications in the entire defense range, to obtain data related to network security, and at the same time to save the data uniformly and generate a secure database [8]. The second stage is to use data mining technology to discover security incidents, analyze potential threats, and predict unknown threats. The third stage is to estimate and perceive the network security situation based on data mining. The overall framework is described in Figure 1.

1.2 Extracting features of network security situation perception

First, it should be made clear that every cyber security information can be stored during the entire process of recognizing the cyber security situation. Considering that gaining knowledge of network security

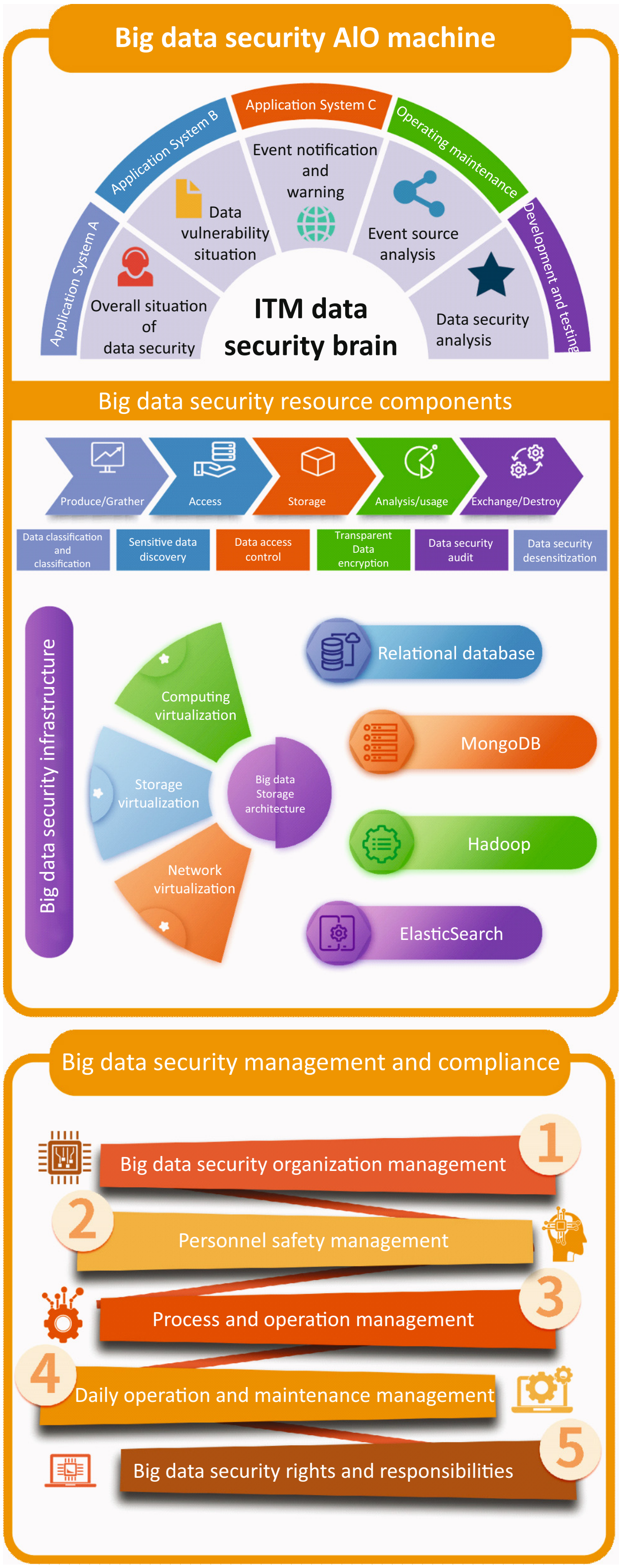
Big data security AIO machine
Application System A
Application System B
Application System C
Operating maintenance
Development and testing
Overall situation of data security
Data vulnerability situation
Event notification and warning
Event source analysis
Data security analysis
ITM data security brain
Big data security resource components
Produce/Grather
Access
Storage
Analysis/usage
Exchange/Destroy
Data classification and classification
Sensitive data discovery
Data access control
Transparent Data encryption
Data security audit
Data security desensitization
Big data security infrastructure
Computing virtualization
Storage virtualization
Network virtualization
Big data Storage architecture
Relational database
MongoDB
Hadoop
ElasticSearch
Big data security management and compliance
1
Big data security organization management
2
Personnel safety management
3
Process and operation management
4
Daily operation and maintenance management
5
Big data security rights and responsibilities

Figure 1: Framework of network security situation perception.

situations is the easiest way to sense network security, this study breaks down network security situation knowledge into network security coefficients and derives the product of network security situation knowledge from a single network node. The nature of the knowledge of the network security situation is obtained by estimating the stability of the sub-channel and the channel through data mining, as shown in Table 1.

Table 1: Features of network security situation perception based on data mining

Network node	Features of network security situation perception	Stability of sub-channel data mining	Stability of channel data mining
1	0101	0.0287	0.0378
2	010100	0.1820	0.0124
3	0101010	0.8256	0.0875
4	01001100	0.2795	0.0074

It can be seen from Table 1 that the features of network security situation perception are directly related to the degree of stability of channel perception, and indirectly related to the degree of stability of sub-channel data mining.

2 Literature review

In the large-scale network environment, the network system is more complex, the diversity of users, business, and system openness become the main characteristics, so that the traditional information security problems become more serious. Based on the above reasons, researchers began to study the network security situation awareness technology, mainly to monitor the current network security situation in real time, find out some potential and malicious network behaviors as much as possible, and provide timely response strategies before they cause bad effects on the network. Trojanowska proposed a comparative analysis of using data mining technology and feature selection software. The most important thing in network security situation perception is to collect and analyze large-scale network security incidents, and at the same time, to effectively describe the information, so that network security managers can quickly grasp the network security situation [9]. However, in practical applications, network security situation perception faces many difficulties: (1) At this stage, the amount of network security alarms is large, the false alarm rate is high, the scale of data to be analyzed is large, and the continuous working time is very long. And the amount of network alarms can generally reach the order of G, of which about 90% of the alarm information is false. Chaudhuri et al. proposed a predictive machine learning method for data mining in the complex problem solving process [10]. Therefore, how to dig out useful information from large-scale data is an important problem that needs to be solved for network security situation perception. (2) Zae et al. proposed a data mining-based overall analysis method of urban road parking lots [11]. Ghosh et al. found that in the process of network security situation perception, data are the basis of research, and the choice of data type plays an important role in the final study results. Data collection includes three parts: device log collection, sensor collection, and data preprocessing. Device logs mainly collect log data generated by different security devices on the network. Sensors can provide more complete data and improve the reliability of analysis results. Pre-processing mainly preprocesses the collected data and deletes some invalid data to obtain more accurate analysis results [12]. Strand, M found that before pre-processing the data, each received network security event is transformed into a standard format that can be processed, which are of many types and large in scale [13].

It can be described by the form of the following tuples:

$$W_i = T_d, \mathrm{ET}, a_i, IP_s, IP_d, p_s, p_d, p_i, S_i, C_i, S_y, \text{other}, \tag{1}$$

where T_d is used to describe the occurrence time of network security events, a_i is used to describe the level of network security event, ET is used to describe the type of network security event, IP_s and IP_d are used to describe the source and destination address of the network security event, P_s and P_d are used to describe the source port and destination port of a network security event, respectively. P_i is used to describe the protocol type, S_i is used to describe the sensor that collected the security event, C_i is used to describe the credibility of the occurrence of the network security event, S_y is used to describe the severity level of the network security event, and "other" is used to describe the remaining information of the network security event. There are differences both in the sources of network security events, and in the format of collected information. So the events need to be converted into a unified form for subsequent processing [14–16].

According to the above-designed experiment, 7 sets of experimental data are collected, and the operation dimension situation index under the two perception technologies is compared. The former is the experimental group and the latter is the control group. In order to more intuitively reflect the differences between the two perception technologies, the results are displayed in the form of a curve graph in the comprehensive situation perception software, as shown in Figure 2.

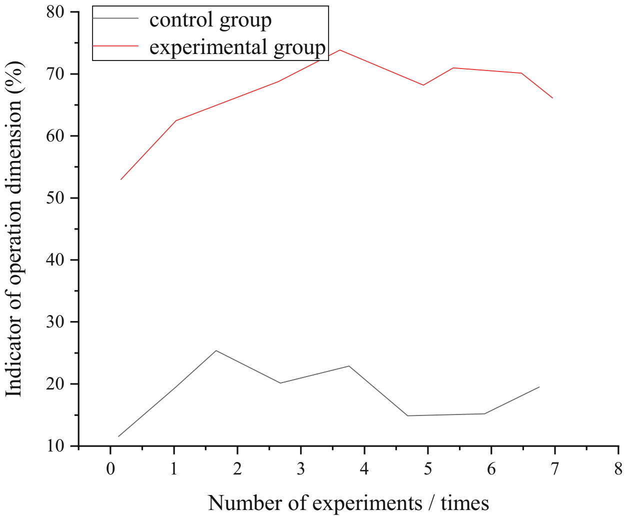

Figure 2: Comparison of the situation index of the perception operation dimension.

The most important part of network security situational awareness is to collect and analyze large-scale network security events and describe the information effectively so that network security managers can quickly grasp the network security situation. However, in practical application, network security situational awareness is faced with many difficulties: at present, with a large number of network security alarms, high false positive rate, large scale of data to be analyzed, and continuous work, the number of network alarms generally can reach the order of magnitude of G, and about 90% of the alarm information is false positive information. Therefore, how to mine useful information from large-scale data is an important problem to be solved for network security situation awareness. The attack activities in the network will generate large-scale trivial alarm information to a large extent, and the correlation is difficult to determine. How to accurately identify the attack behavior is a problem that security situation awareness needs to solve.

3 Methods

3.1 Extraction of association rules based on data mining

Through data mining methods, pattern mining, pattern analysis, and learning are performed on network security event datasets, so as to complete the extraction of network security situation rules and provide a basis for network security situation assessment. Assume that W represents the set of all project elements, which can be described as $W = \{W_1,W_2,...,W_n\}$. The dataset $R = \{r_1,r_2,...,r_n,.,\}$, where the element r_i of R is a set composed of elements in the set W, that is, $r_i \subseteq w$.

Definition 1. Set C is composed of elements in R. If the number of data in set R matches $C \subseteq r_i$ is 1, then the support level of set C in dataset R can be calculated by $\sup(C) = l/n$. If $\sup(C)$ exceeds the minimum support level threshold, then the set C is the frequent k-item set of the dataset R.

Definition 2. If the set C and D accord with $A \subseteq W \cap D \subseteq W$, then the confidence level of $C \to D$ can be described as $\frac{\sup(C \cup D)}{\sup C}$. Association rules are mainly to mine the $C \to D$ that meets the minimum support and minimum confidence in the dataset. It is proposed based on the shopping basket problem. Through the mining of the frequent item sets of the thing set, the association rules between things are discovered. It has been widely used in many fields such as retail, finance, and e-commerce [17–19]. Association rule mining mainly includes two steps. First, mining frequent item sets that meet the minimum support level, and second mining association rules that meet the minimum confidence level according to frequent item sets. In view of the large scale of network security situation data, association rule mining is carried out with the help of Hadoop platform [20]. The subset of item sets is obtained through Map function, and the support level of all subsets is collected in the instrument through Reduce function, so as to obtain the support level of frequent items and mine the frequent item sets in the dataset. The detailed steps are as follows:

Input: The input path of the original network security dataset R, the minimum support level β.
Output: Frequent item set files meeting the minimum support level.

(1) According to the path of the input file, divide the original network security dataset from the frequent item set file with minimum support into n data subsets of uniform size, and format each row of the subset to form <key, value> pair, key represents character offset, while value represents data information;
(2) Use the Map function to read each <key, value> pair in the subset, and use the split function to resolve the value into the set;
(3) Treat all subsets as output key, and assume that the value of each subset is 1;
(4) Call the optional Combine function. In the massive network security data, all Map terminals will form large-scale key–value pairs with the same key value. If all the key–value pairs obtained are transmitted to Reduce using the network, efficiency will be greatly reduced. Therefore, the Combine function is used to merge the same key–value pairs together [21].

Reduce task

(1) Sort the key–value pairs transmitted by the Combine function, merge the key–value pairs with the same key value to get <key, L(value)>. Read it through the Reduce function, and values in L(value) can be accumulated to obtain the support number of the key set in the network security dataset R, that is, the global support level of the frequent candidate item set is obtained on the Reduce terminal;
(2) Based on the minimum support level, the candidate item set that exceeds the minimum support level and the corresponding support level are transmitted to the external table of the data storage framework. It is used to query frequent cameras by using the external table, and also can be regarded as the input or output file of the Map Reduce program. Mining the association rules of the network security situation is to mine the elements in the frequent item set that meet the minimum confidence level. The detailed process of mining association rules is as follows:

Input: the minimum confidence level δ.
Output: the association rule that meets the minimum confidence level δ.

(1) Use the Map function to call the “setup” method, and use the “setup” method to get the database connection;

(2) Divide the frequent item set in the external table of the data storage framework into n data subsets of uniform size, complete the formatting process, and resolve each row of data into <key, value> pairs
(3) Analyze the elements in the frequent item set field in "value" to obtain the value of C, D, S Value, and save C, D to the set;
(4) If the confidence level exceeds the predetermined value, then the association rules between the subset and the frequent item set except for the subset are the association rules. The subset and the difference set form the key value, and the confidence level is "value."

3.2 Assessment of network security situation

The network security situation refers to the distribution of intrusion in the monitored network and its impact on security objectives over a period of time. Network security situation information is mainly related to time and space. For an independent node, after an attack, the attack index and resource impact level will change. For the entire network, after an intrusion, the attack focus distribution will also change.

When analyzing the network security situation at a certain moment, it is necessary to accumulate the risk value of the assessment time window. As time goes by, some alarm events slowly leave the time window and are filled by new alarm events. The frequency of alarm events can reflect the degree of network intrusion.

The security risk level of the network node is calculated according to the alarm time after merging, which is mainly related to the alarm confidence level c, the alarm severity level v, and the resource impact level h. The alarm confidence level is calculated by the initial definition and merging, the alarm severity level is set, and the resource impact level is related to the network configuration and service.

In addition to the above analysis, the security defense level e_n of the node and the alarm recovery coefficient S_n need to be considered.

The security situation assessment value of an independent node can be obtained by the following formula:

$$Z_n(t) = \sum c_i v_i h_i / e_n s_n. \tag{2}$$

In the process of assessing the network security situation, it is necessary to analyze the risks of different nodes in the network. Because the positions and functions of the nodes in the network are different, the key levels of the nodes are also very different, so the weight W_n of the nodes need to be calculated.

The network security risk value can be calculated by the following formula:

$$Z_N(t) = \sum w_n Z_n(t). \tag{3}$$

In the above analysis, the network security alarm target node is mainly analyzed, but in practical applications, the activity level of the intruder still needs to be analyzed in the complete security situation. Therefore, it is also necessary to obtain the attack index of the intruder according to the network security alarm source address. The formula is described as follows:

$$Q_n(t) = \sum c_i v_i. \tag{4}$$

Through the above analysis, the change curve of the network security situation assessment value over time can be obtained. Based on the network security situation assessment value, the prediction of the network security situation can be realized, and the perception of the network security situation can be completed.

4 Results

4.1 Experimental data

The experimental time window is 1 min. There are about 1,000 time windows in the attack detection dataset within 16 h. The experimental data time is from 8 o'clock to 0 o'clock, a total of 16 h. 39,125 pieces of data are obtained through one-way filtering. There are seven security events in the entire dataset, with a total of three types of attacks. The specific network security events information is described in Table 2.

Table 2: Security event table

Types	Time	Intruded party	Duration (s)
U2R	8:25:12	168.12.115.50	242
R2L	8:56:35	168.12.110.50	3
PROBE	9:51:22	168.12110.50	41
U2R	10:29:05	168.12.112.50	182
R2L	11:50:29	168.12.100.100	15
U2R	12:35:51	168.12.114.50	86
PROBE	21:25:13	168.12.1132.50	18

4.2 Experimental process and analysis

In the experimental environment, 168.12.xx represents the internal network. In order to obtain complete intrusion data, no defense was performed during the experiment. Therefore, when the network security situation is perceived, the method of artificially changing the experimental data is also used. After a certain period of time, simulate the changes in network security situation perception when false alarm or missing alarm or security defense works, so as to verify the effectiveness of the method in this study [22]. Figures 3 and 4 adopt the method proposed in this study to calculate attack index of the network security situation perception from the perspective of the original situation and artificial changes.

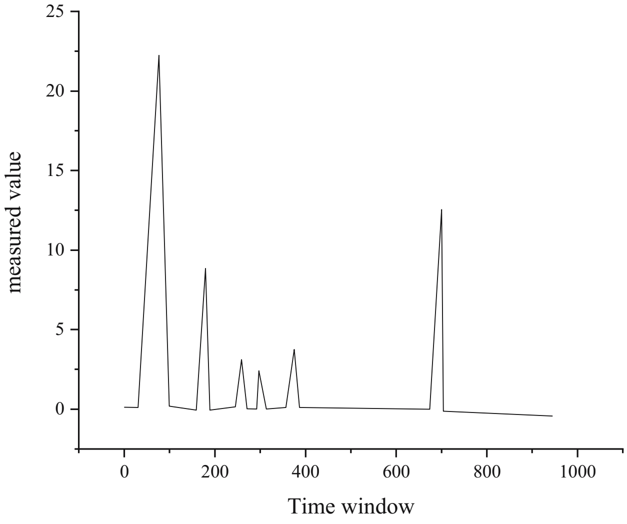

Figure 3: Change graph of original data attack index.

In Figure 3, the most primitive security threats are shown in the range of 0–500 windows. Since the network has no security defenses, both accessed from outside the network and related to network security events enter the subnet, so it can be reflected in the attack. After the 500th time window, defensive processing for network security events begins to work.

Figure 4 shows the attack index obtained in this study after manually reducing network security events of the 295th time window. It is found that after the security events are cancelled, the corresponding window

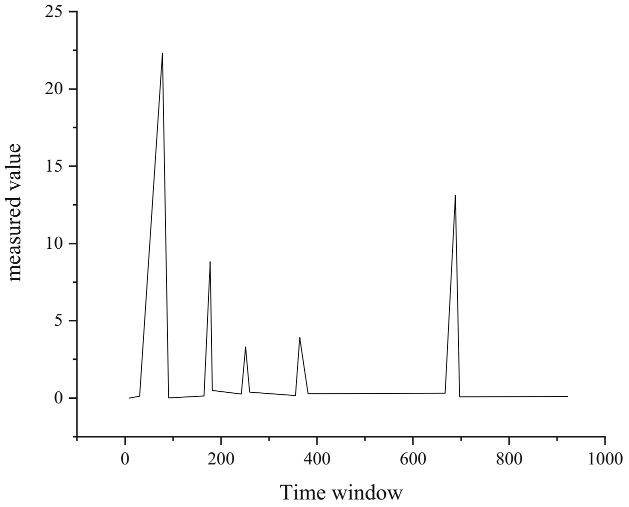

Figure 4: Change graph of attack index after scanning is reduced.

attack index is reduced to 0, indicating that the method in this study can effectively realize the network security situation perception. The above analysis is the perception of changes in the network security situation with the method in this study. The following analyses the actual changes in the network security situation. Described by the defense index, it is analyzed whether the perception results of the method in this study are consistent with the actual changes in the network situation.

In Figure 5, network security events appear from the 20th time window, and then the defense index decreased until the 500th window. This is mainly due to the lack of network security defenses during this period of time. In 500th window, the defense index increases, which is consistent with the perception results of the method in this study, indicating that the method in this study can perceive changes in the network security situation.

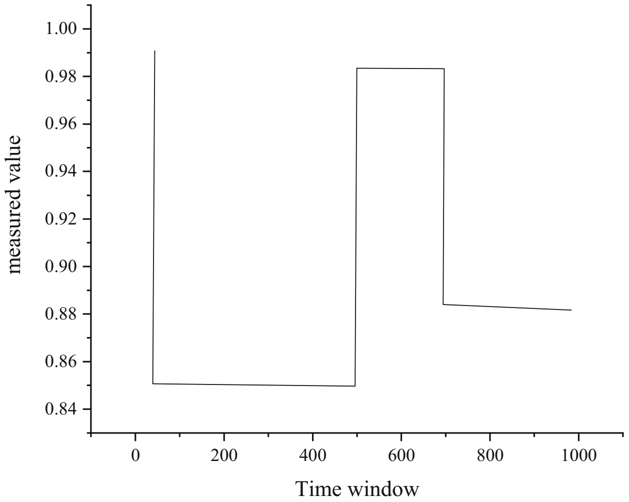

Figure 5: Change diagram of the original data defense.

In Figure 6, starting from the first attacked window, the defense index begins to decrease. In order to simulate the added network defense, the 295th time window of network security events is reduced from the original data, and defense index has increased significantly, which is consistent with the perception results

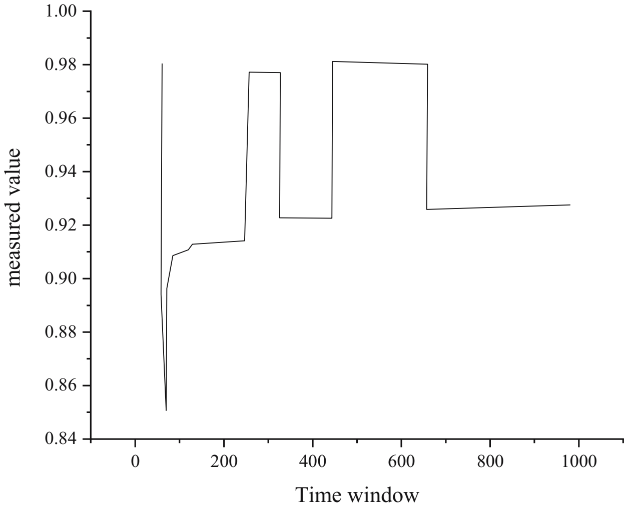

Figure 6: The defense situation of a simulated network security event being intercepted.

of the method in this study, further verifying the effectiveness and reliability of the method in this study in the perception of network security events.

5 Conclusion

Data mining-based network security knowledge technology studies have shown that the accuracy of the concept of the security situation of the designed technology is much higher than that of traditional technology, and that the designed concept technology can implement the real concept of network security, network security situation. Therefore, data mining-based network security status recognition technology is the most effective and reliable way to recognize network security status. Based on the overall framework design, the network security situation is modeled and evaluated. It can provide a simple and effective overview of the network security situation and give a basic idea of the relevant issues. Simplified processing of a database of major network security events through data mining technology is achieved through the constant extraction of network security status connection rules. A network security situation model is established in accordance with the rules of the association, and the network security situation is calculated and predicted. Accuracy of knowledge about the network security situation is an important tool to ensure the sustainable operation of network security. It is reasonable to consider data mining as a basic technology for knowledge of network security situations.

Conflict of interest: The authors declare that they have no competing interests.

Data availability statement: The data used to support the findings of this study are available from the corresponding author upon request.

References

[1] Shirono T, Niizeki T, Iwamoto H, Shimose S, Torimura T. Therapeutic outcomes and prognostic factors of unresectable intrahepatic cholangiocarcinoma: A data mining analysis. J Clin Med. 2021;10(5):987.

[2] Salazar-Carrillo J, Torres-Ruiz M, Davis CA, Quintero R, Guzmán G. Traffic congestion analysis based on a web-GIS and data mining of traffic events from twitter. Sensors. 2021;21(9):2964.

[3] Nelson LK. Knowledge discovery in the social sciences: A data mining approach. Contemporary Sociol. 2021;50(4):346–8.

[4] Dudurych IM. The impact of renewables on operational security: Operating power systems that have extremely high penetrations of nonsynchronous renewable sources. IEEE Power Energy Mag. 2021;19(2):37–45.

[5] Latif A, Fitriana LA, Firdaus MR. Comparative analysis of software effort estimation using data mining technique and feature selection. JITK (J Ilmu Pengetah dan Teknologi Komput). 2021;6(2):167–74.

[6] Peji A, Molcer PS. Predictive machine learning approach for complex problem solving process data mining. Acta Polytechnica Hungarica. 2021;18(1):45–63.

[7] Dai X, Weng W, Chen G, He J, Shen J. An overall analysis method of urban road parking lots based on data mining. Int J Security Netw. 2021;16(2):105.

[8] Ceresnak R, Kvet M, Matiasko K. Increasing security of database during car monitoring. Transportation Res Procedia. 2021;55(11):118–25.

[9] Trojanowska BK. Women's rights facing hypermasculinist leadership: implementing the women, peace and security agenda under a populist-nationalist regime. Feminist Leg Stud. 2021;29(2):231–49.

[10] Chaudhuri S, Roy M, Mcdonald LM, Emendack Y. Coping behaviours and the concept of time poverty: A review of perceived social and health outcomes of food insecurity on women and children. Food Security. 2021;13(4):1049–68.

[11] Zae D, Pietro M, Reali L, Waure CD, Ricciardi W. Prevalence, socio-economic predictors and health correlates of food insecurity among Italian children-findings from a cross-sectional study. Food Security. 2021;13(1):13–24.

[12] Ghosh S, Brooks B, Ranmuthugala D, Bowles M. Investigating the correlation between students' perception of authenticity in assessment and their academic achievement in the associated assessment tasks. J Navigation. 2021;74(2):293–310.

[13] Strand M, Fredlund P, Boldemann C, Lager A. Body image perception, smoking, alcohol use, indoor tanning, and disordered eating in young and middle-aged adults: findings from a large population-based swedish study. BMC Public Health. 2021;21(1):1–12.

[14] Qin T, Cook M, Courtney M. Exploring chemistry with wireless, pc-less portable virtual reality laboratories. J Chem Educ. 2021;98(2):521–9.

[15] Bocksnick J, Napier J. Re: risky business: doctors' understanding of statistics – are we following a naked emperor? BMJ. 2021;73(7):40–3.

[16] Surya T, Dewi C, Hendijani RB. Key decision-making factors of moocs users towards paid moocs. Int J Educ Econ Dev. 2021;12(2):151.

[17] Gurariy G, Randall R, Greenberg AS. Manipulation of low-level features modulates grouping strength of auditory objects. Psychological Res. 2021;85(6):2256–70.

[18] Ibrahim M, Abdul-Halim S, Ishak MY, Hassan S. The local community awareness on Langkawi unesco global geopark status: case of Kampung Padang Puteh, Langkawi, Malaysia. Int J Geoheritage Park. 2021;9(1):233–41.

[19] Cuevas E, Becerra H, Luque A, Elaziz MA. Fast multi-feature image segmentation. Appl Math Model. 2021;90(5):742–57.

[20] Yu Y, Lu J, Shen D, Chen B. Research on real estate pricing methods based on data mining and machine learning. Neural Comput Appl. 2021;33(9):3925–37.

[21] Li H. Time works well: dynamic time warping based on time weighting for time series data mining. Inf Sci. 2021;547:592–608.

[22] Sharma A. An optimal routing scheme for critical healthcare HTH services – an IOT perspective. International Conference on Image Information Processing; 2017. p. 1–5.