Hindawi
Scientific Programming
Volume 2022, Article ID 4960360, 13 pages
https://doi.org/10.1155/2022/4960360

Research Article

Big Data-Driven Hierarchical Local Area Network Security Risk Event Prediction Algorithm

Wei Zhou

Information Engineering Department, Suzhou University, Suzhou 234000, China

Correspondence should be addressed to Wei Zhou; weizhou@ahszu.edu.cn

Received 20 December 2021; Accepted 18 January 2022; Published 8 February 2022

Academic Editor: Sheng Bin

Big data processing technology has attracted a lot of attention due to its forecasting and warning of Internet security situation. The current risk assessment system still has problems such as high false alarm rate and excessive reliance on expert knowledge in the security defense system. Based on the big data-driven principle, this paper constructs a hierarchical local area network security risk event prediction model and proposes a predictive complex event processing method. The model building process is evolved and improved on the basis of the scoring function. The establishment method of vulnerability database and vulnerability association database is introduced in detail. At the same time, the problem of the difference between the structure and identification method of the information in the information database and the vulnerability database is solved, and the effect of timely modification when the data do not match is realized. Experimental results show that the algorithm has an accuracy of 98.75% and a fault tolerance rate of 0.0035, which promotes the accuracy of the network risk assessment results based on multistage network attacks.

1. Introduction

In the era of big data, the Internet, sensor networks, social networks, etc. continue to generate a large amount of data. When facing a large amount of information, the shorter the time it takes to make decisions [1]. In practical applications, such as intelligent transportation, medical and health, and financial risks, users hope to dynamically process the complex events identified from the original event stream and summarize the most significant and most relevant parts of them and be able to effectively infer the predecessor or successor events that may occur when a given event occurs. However, the current predictive analysis methods for complex event processing are not mature enough, and there are still many challenges in the field of streaming big data, and more in-depth and specific research is needed [2–5]. In order to solve the current challenges and problems, this article uses hierarchical local area network to conduct in-depth study of complex event processing. Hierarchical local area network is a reasoning model based on a solid mathematical theory, which can efficiently perform uncertain knowledge. Aiming at the problem of the concept drift of data distribution, this paper proposes different model construction methods for data gradual change and data sudden change, so that it can better meet the real-time changes of data distribution in the real environment and achieve the best prediction. The applications of big data processing are mainly divided into two categories: batch processing and streaming big data processing. Batch processing, as a traditional big data processing mode, is oriented to a large amount of historical data, and the response time is generally longer. The advantage is that it can process large-scale historical data, but it cannot meet the needs of real-time response. Streaming big data processing is oriented to high-speed data streams, which are processed at one time, and the response time is greatly shortened, which can meet real-time requirements but the scale of processed data is relatively limited. At present, with the continuous improvement and deepening of big data processing technology, streaming big data processing technology has received more and more attention and has been widely used in transportation, finance, telecommunications, and other fields [6–8].

Some domestic and foreign researchers have studied predictive analysis for many years and have proposed a series of models and algorithms. Fathi et al. [9] proposed a predictive analysis method based on a deep belief network.In the work of Greco et al. [10], a stack autoencoder model is used to learn general event stream features and trained in a greedy hierarchical manner. Dautov et al. [11] believed that the predictive analysis method using deep learning models can usually achieve better accuracy. Abkenar et al. [12] used a probabilistic event processing network to detect complex events and then used these complex events to train a multilayer hierarchical local model to predict future events. Their prediction model uses an expectation maximization (EM) algorithm [13–15]. Ghorbanian et al. [16] proposed a framework to incorporate predictive analysis technology into CEP applications. They extended CEP solutions with predictive capabilities, defined the key aspects of the combination of these technologies, and summarized how CEP and predictive analysis can benefit from the joint solution. Lana et al. also proposed a conceptual framework for predicting CEP by combining CEP and predictive analysis [17]. Researchers proposed an active architecture that uses a combination of machine learning and CEP to predict historical data [18] and proposed a method called adaptive moving window regression (AMWR) for dynamic IoT data,that is, adaptive prediction algorithm, which is evaluated with real-world use cases. Some scholars have proposed a basic framework that combines time series forecasting and CEP to monitor product quality to ensure its quality throughout the supply chain cycle, summarized the research progress of machine learning in the field of text classification [19–21], discussed possible solutions to common problems encountered in processing text classification, and affirmed the future development of this technology. In order to solve the problem of overlapping of software vulnerability classification, the researchers performed text clustering on the vulnerability description field of the vulnerability database NVD (national vulnerability database) and clustered 40,000 vulnerability data in the NVD into 45 types of typical vulnerabilities [22]. The HT-SVM multiclassifier based on the vulnerability distribution is constructed to improve the classification effect [23]. In order to extract effective vulnerability feature words, the use of CHI for feature extraction of vulnerability text was proposed, and a binary tree SVM category prediction model based on category entropy was constructed, which improved the accuracy of prediction [24–26].

This paper mainly studies the technical problems of complex event processing in the streaming big data environment. Based on the Bayesian network model, a complex event processing system framework and different processing methods are proposed according to different data distributions. In recent years, with the continuous growth of the network scale, the current large-scale network nodes are becoming larger and larger, and the structure is becoming more and more complex. A large number of network devices (systems) are deployed in the network, and events are generated by various devices (systems). The format is different, and the event risk classification standards are also very different. Due to the lack of automatic and comprehensive analysis methods for massive event information, it is impossible to quickly extract, locate, and summarize important security events, and it is difficult to effectively evaluate the current network security situation and implement. In view of the current challenges and problems, the research content of this article is proposed. It is mainly based on the theoretical knowledge of hierarchical local area network structure learning, integrating the characteristics of event flow and structure learning, and constructing a hierarchical local area network model based on the characteristics of streaming data. After that, the structure and parameters of the constructed model are optimized, and finally a series of inferences and predictions are made on related events on the basis of the constructed event network. This paper provides a detailed analysis and explanation of the basic theories of complex event processing related knowledge and a specific introduction and comparative analysis of the current more mature complex event detection models and summarizes their respective advantages and disadvantages, and at the same time, different event context types are introduced and summarized. The author describes the basic theory and definition of hierarchical local area network, as well as the learning method of hierarchical local area network, mainly including the structure learning method and parameter learning method of hierarchical local area network, and on the basis of dynamic hierarchical local area network, a streaming data prediction Bayesian network model is proposed.

2. Big Data-Driven Hierarchical Local Area Network Distribution

2.1. Big Data-Driven Topology. Data-driven is to extract or mine knowledge from a large amount of data. Data mining can be considered as the process of knowledge discovery in a database, which is divided into data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, and knowledge representation. ETL is the abbreviation of data extraction, transformation, and loading, and it is an important part of the realization of data warehouse. ETL extracts data from dispersed heterogeneous data sources and transfers into a temporary space for cleaning, conversion, and integration and finally saves the data in a data warehouse as the basic data for online analysis and processing or data mining.

$$F(x, t | t \in R(c, n)) = (a_1, a_2, \cdots, a_n) \times (a_1, a_2, \cdots, a_n)^T. \quad (1)$$

When new data arrive, the above formula can be used to update the density index to update the cluster. In online clustering, the cluster centers and clusters obtained by offline clustering are directly used for division, avoiding the need to relearn each time and updating the clusters based on the new data.

If a new class is generated during online clustering and the number of samples divided into the new class is large, the new class will be learned to obtain the structure and parameters of the corresponding hierarchical local area network and uses them for real-time prediction. If real-time prediction is performed when the current event context is close to multiple existing clusters, a combination method of multiple hierarchical local models is used to predict.

$$\sum_{i=1}^{n} A_i - \frac{\sum_{i=1}^{n} \left(a_1, a_2, \cdots, a_n\right)^T}{\sum_{i=1}^{n} \left(b_1, b_2, \cdots, b_n\right)^T} = 0. \tag{2}$$

File transfer protocols can be realized mainly in two ways: FTP protocol and HTTP protocol. FTP realizes the two-way control of file transfer. It must rely on an application program running FTP service. The user connects to other servers running FTP protocol through the application program. The HTTP protocol is a hypertext transfer protocol, which implements hypertext transfer from a server that provides the World Wide Web to a client that accesses the server through a browser.

$$\iint \left(c_i - d_i\right) \times B_i dcdi - \iint \left(c_i - d_i\right) xdx = 0. \tag{3}$$

The computer can send files in the form of hypertext and save them to a designated location on the server. To use the FTP protocol to transfer files, you must run an FTP application or browser plug-in on the client. Since the FTP protocol transfer is stable and fast, it can easily realize the resumable transfer.

2.2. Local Area Network Connection. All nodes in the local area network node set V are connected to their candidate parent nodes one by one. If the newly added edge increases the scoring function value, the newly added edge will be saved in the edge set set *E*, and then the new edge *E* will be saved during learning in the contraction stage. If the scoring function value is not reduced by deleting the edge, remove the edge from *E*, and use EM algorithm to learn the parameters of Gaussian mixture model. When the parameters need to be updated, first use the current *m* parameter to calculate the distribution of the hidden variables for the changed samples, and then use the maximum likelihood based on the calculated distribution of the hidden variables law.

$$\lim_{x \longrightarrow \infty} \log \frac{xe_i}{f_i} - \lim_{x \longrightarrow \infty} \log \left(xe_i\right) - \lim_{x \longrightarrow \infty} \log \left(f_i\right) = 0. \tag{4}$$

To achieve mail filtering between internal and external networks to prevent internal mail from directly flowing into computers on the external network, it is necessary to establish a mail front-end processor between the internal and external networks to filter the sending of mail according to the mail address through the mail policy. Mail transmission between the internal and external networks cannot be carried out directly.

It needs to be filtered by the firewall between the gatekeeper and the internal and external networks to prevent malicious attacks on the mail server. Each member unit is in a unified wide area network, and firewalls are set up between each local area network to prevent a certain local area network from being attacked, and all distributed mail systems are paralyzed.

$$\sum_{i=1}^{n} \left(c_i - d_i\right) \log xe_i = \sum_{i=1}^{n} f(x) \left[\exp(n) - \exp(x)\right] = 0. \tag{5}$$

Through the Web service interface, the mail server's mail sending record, attachment name, attachment size, and other information are synchronized to the intermediate database at regular intervals, and these data are synchronized to the data warehouse during the ETL process of the data analysis in Table 1, so as to realize the sending, blocking, and analysis of spam filtering.

The last column represents the overall detection rate of the four types of attacks: DoS, Probe, R2L, and U2R, that is, the abnormal detection rate in the two classifications. Compared with other methods, the method in this article has a better classification effect in terms of both the overall and the partial. Compared with all the features, although the detection rate of the normal class and the DoS class has decreased, it also retains the basic accuracy rate. In addition, the Probe, R2L, and U2R types with lower detection rates have also been significantly improved. Therefore, the features selected by the method in this paper effectively remove the redundant features of the original data set, retain the basic relevant features, and reduce the dimensionality while maintaining the accuracy of the classifier.

2.3. Network Hierarchy. In order to verify the effectiveness of the proposed network hierarchical model, preliminary experiments were carried out using the LLDOS1.0 data set of MIT Lincoln Laboratory. According to the results of the experimental data set, one can restore a set of 6 hosts: 131.84.1.31 (Web server), 172.16.115.20 (Solaris), 172.16.115.50 (Solaris), 172.16.115.10 (Solaris), 172.16.115.11 (Windows), and a network composed of 172.16.115.12 (Windows). Whether it is Algorithm 1 or Algorithm 2 after a finite number of iterations, it can be seen that the attacker's attack range or utility affects the entire network, which is consistent with the fact that the network can be attacked.

$$\frac{\sum_{i=1}^{n} \left(g_i^{(j)}\right)^{1/l}}{\sum_{i=1}^{n} (t)^{1/l}} - \frac{\sum_{i=1}^{n} \left(f_i^{(j)}\right)^{1/l}}{\sum_{i=1}^{n} (t-1)^{1/l}} = 0. \tag{6}$$

In terms of defense measures, the conclusions drawn by Algorithm 1 and Algorithm 2 are consistent with each other for repairing the Sadmind node on the 103 172.16.115.20 Solaris server and the ICMP node on the 131.84.1.31 Web server. This is because the value of the empirical probability tends to be more realistic in the simulation environment, so the results obtained are similar, which also proves the correctness of the model in Figure 1 from the side.

Hardware access control needs to monitor the hardware status of each client computer, discover the client's access device in time, and determine whether it has access permissions. Hardware access needs to record the user identity of the client computer and monitor the client computer in real time or regularly, so the C/S structure is adopted. The hardware management in this subject reads the security management data generated by the UniAccess network access control system provided by Liansoft Technology,

Table 1: Analysis of network filtering.

Detection rate	Data network 1	Data network 2	Data network 3	Data network 4
RL 1	−0.33	−0.94	0.35	7.18
RL 2	0.93	−0.37	-2.51	6.77
RL 3	0.72	0.69	1.05	6.89
RL 4	0.30	0.95	0.32	7.42
RL 5	5.07	0.44	0.90	0.49
RL 6	5.92	0.89	0.46	1.92

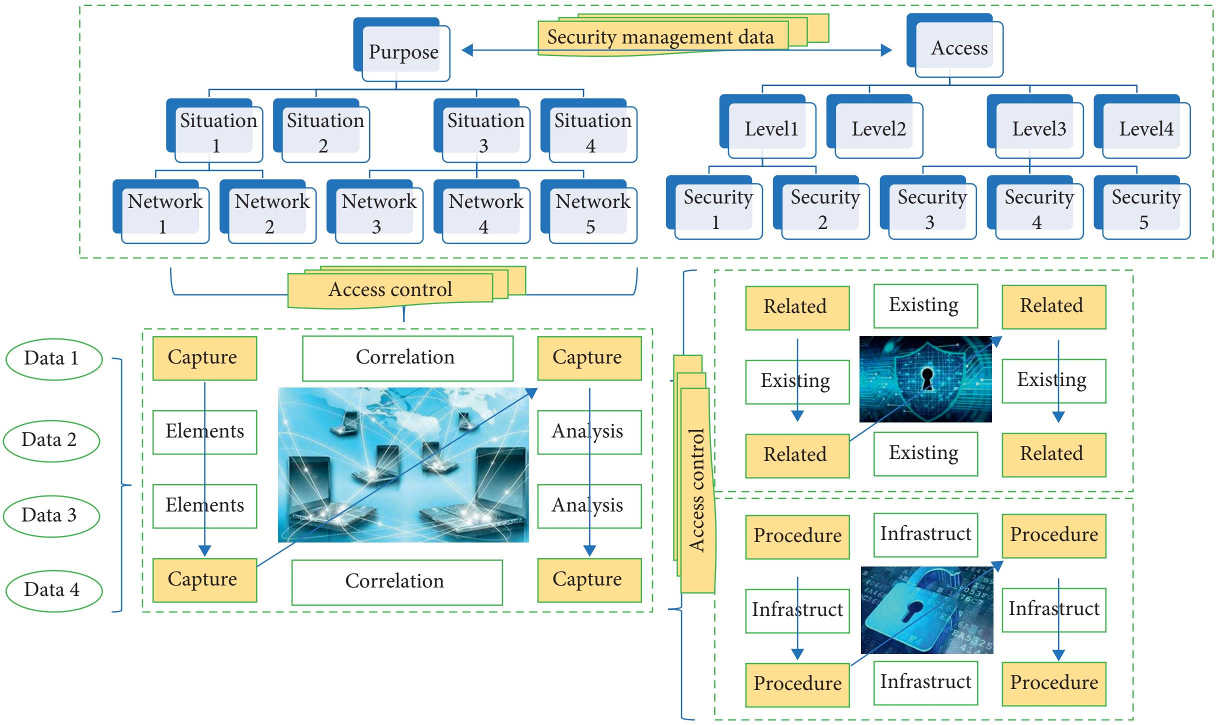

Figure 1: Network hierarchy topology.

which is being used by National Nuclear Power, to monitor and track the data transmission through hardware means such as mobile media.

$$\left\langle \frac{\left(g_i^{(j)}\right)^{1/l}}{\sum_{i=1}^{n}\left(g_i^{(j)}\right)^{1/l}} \middle| \frac{\left(g_i^{(j)}\right)^{1/l} - \left(f_i^{(j)}\right)^{1/l}}{\sum_{i=1}^{n}\left(g_i^{(j)}\right)^{1/l}\left(f_i^{(j)}\right)^{1/l}} \right\rangle = \left\langle \begin{matrix} 1 \\ 1 \end{matrix} \middle| \begin{matrix} -1 \\ 0 \end{matrix} \right\rangle. \tag{7}$$

The file audit management needs to realize the audit of the content of the external files through the workflow. There is no large amount of real-time data calculation, so the B/S structure is adopted to build the system. Centralized data management is to manage sensitive data generated by other systems. It only needs to realize the synchronization of the data layer from the server side to the server side. The database server in the data center provides users with the required processed data results according to user needs, as well as B/S structure data report and analysis tool.

$$\lim_{i,j\longrightarrow\infty} C_i\left[g_i^{(j)}, f_i^{(j)}\right] - \frac{\sum_{i=1}^{n}\left(g_i^{(j)}\right)^{1/l}}{f_i^{(j)}} \times \sum_{i=1}^{n}\left(h_i^{(j)}\right)^{1/l}, for\left[f_i^{(j)} > 0\right]. \tag{8}$$

The abstract definition of the optimal solution of the attack network is given, the probability distribution is used to describe the uncertainty of the strategy set, and then the defensive subset mining algorithm of the polynomial time uncertainty attack strategy set and the known attack strategy are given for a set of algorithm for solving random defense measures.

2.4. Network Safety Factor. Network security risk assessment refers to the assessment of the vulnerability of the network system, the threats it faces, and the impact and loss caused by the use of the vulnerability and the identification of the security risk of the network system according to the possibility of security incidents and the degree of loss. Security management personnel grasp the current network risk situation, and based on the assessment results, network security administrators provide detailed and reliable security analysis reports and vulnerability repair recommendations, so as to take early measures to protect the security of various devices and data in the network. It specifically includes risk assessment, security strategy, plan design, implementation of security elements, etc.

$$\oiint \lim_{x\longrightarrow\infty} C(\pi i, i-1)D_i - \frac{p_i}{\sum_{i=1}^{n} o_i} = 0, \ (i = 1, 2, \cdots, m). \tag{9}$$

Document outsourcing audit adopts JAVA language and realizes the information management system of B/S structure. The system adopts the layered design mode of MVC under the J2EE specification. The system is divided into presentation layer, interaction layer, application layer, domain layer, and data access layer. However, strict layering and complete interface modes are not advocated during system development.

It mainly includes the identification and assignment of risk assessment elements, that is, to identify information assets and assign values to analyze threats and assign values to the possibility of threats, identify the vulnerability of information assets and determine the severity of the vulnerability, calculate the possibility of a security incident and the loss caused by the threat and vulnerability, and finally calculate the risk value of the information asset based on the importance of the information asset in Table 2.

In the process of data storage, we can learn from the use of substorage technology, data dynamic/static desensitization, homomorphic security encryption, data disaster recovery, and other technologies to improve its storage security. In the process of data processing, we can learn from the use of account management, identity authentication and authorization, access control, data traceability analysis, and other technologies to ensure that data access is manageable and controllable and traceable.

$$\prod \frac{|r_i - s_i|}{|t_i - s_i|} \times \prod E(i, i-1) - \prod \frac{|r_i - s_i|}{|t_i - s_i|} \times f(t, s|r \cup s = C) = 0. \tag{10}$$

The system first collects various events from monitored objects such as network equipment, security equipment, application systems, and operating systems. It uses filtering and merging methods to achieve noise reduction before collection and storage and standardizes heterogeneous event information in a unified format. After warehousing, the original event information database is formed, and then the selected original event information is correlated and analyzed, and important security events are screened out. Through the integration of various modules, the risk assessment situation diagram is displayed, and the security administrator makes the next decision.

3. Big Data-Driven Hierarchical Local Area Network Security Risk Event Prediction Model Construction

3.1. Big Data-Driven System Detection. The basic operation of the big data-driven system protocol is as follows: the SNMP management station first performs setting (SET) and reading (GET, GET NEXT); through these two operations, a certain variable in the MIB library of the SNMP intelligent agent is processed. The SNMP intelligent agent can also take the trap operation (Trap) to actively send abnormal alarms to the SNMP management station and consider network management as a distributed application, which includes the one-to-many relationship between the SNMP management station and the SNMP intelligent agent.

$$\frac{\sum_{i=1}^{n} r_i - s_i/t_i - s_i}{\sum_{i=1}^{n} f(t, s)} - \frac{\sum_{i=1}^{n} u_i - v_i/w_i - v_i}{\sum_{i=1}^{n} f(t-1, s-1)} = 0. \tag{11}$$

SNMP authenticates and develops access strategies through the community mechanism. The SNMP protocol defines the accessible MIB library of network management information objects in a hierarchical and structured form and provides a convenient and fast method for exchanging management information between the SNMP agent and the

Table 2: Materiality calculation information risk description.

Calculation name	Information risk content	Vulnerability value
Access Vector 0	No effect N/Partial effect P/Full effect F	1.05
Access Vector 1	No effect N/Partial effect P/Full effect F	0.32
Access Vector 2	No effect N/Partial effect P/Full effect F	0.49
Access Vector 3	No effect N/Partial effect P/Full effect F	1.92
Access Complexity 1	No effect N/Partial effect P/Full effect F	-2.63
Access Complexity 2	No effect N/Partial effect P/Full effect F	-0.41
Access Complexity 3	No effect N/Partial effect P/Full effect F	0.25
Authentication 1	No effect N/Partial effect P/Full effect F	0.71
Authentication 2	No effect N/Partial effect P/Full effect F	0.76
Authentication 3	No effect N/Partial effect P/Full effect F	0.17

SNMP management terminal. The message is the basic unit of SNMP exchange, and its composition is an external message encapsulation and an internal (PDU).

The premise of using this model for analysis is to abstract the various components in the network, the vulnerability information on the components, and Workshop1 can access the Apache server and Workshop2 can access Table 3. The database server and the working machines can access each other.

In SNMP v2, the format used by the TRAP protocol data unit is the same as that of all other SNMP v2 protocol data units (except GET-BULK), which simplifies the task of the receiver; the system time in the TRAP protocol data unit variable bundle list is the first variable, and the object identifier of Trap is the second variable, followed by the corresponding object instance; SNMP v2 TRAPPDU also has no response mechanism. Venus Star Intrusion Detection System, Rising Network Anti-Virus System, and mainstream routers and switch products all support sending event and log information in SNMP TRAP mode.

3.2. Hierarchical Local Area Network Matching. The article uses a hierarchical local area network linear programming method to solve the problem of random defense measures for known attack strategy sets. The attack strategy set {A} of the attacker AMax and the defense strategy set {D} of the defender DMin are both related to the network scale and network nodes. The number grows exponentially.

$$\partial \frac{|u_i - v_i|}{|w_i - v_i|} / \frac{|u_i - v_i|}{|t|} + \partial \frac{|r_i - s_i|}{|t_i - s_i|} / \frac{|r_i - s_i|}{|s|} = 0. \tag{12}$$

There is no need to enumerate every element in the defense strategy set {D} of the defender DMin and randomly select k strategies from them for calculation. In the process of solving, there is no need to calculate all the utility functions $f(x)$. The number of iterations can ensure the quality of the solution.

Each node v in the network represents a security alarm, and the directed connection r between nodes represents the causal relationship between each alarm. Each node contains a conditional probability matrix representing the relationship between the node and its neighboring nodes, and each element in the matrix represents the conditional probability of the causal relationship between the node and its neighboring nodes. The output layer node Vout of the local area network represents various security events.

$$\left\{\nabla|u_i - v_i|/|u|, \nabla\nabla|u_i - v_i|/|u|, \nabla\nabla\nabla|u_i - v_i|/|u|, \ldots,\right\} \longrightarrow \{\nabla|u - v|/|u|, (u, v) \longrightarrow (0, 0)\}. \tag{13}$$

At time t, the attacker AMax randomly selects a subset of the attack strategy according to a certain probability. For this attack strategy, the defender randomly adopts a set of defense strategy sets for defense. At time $t + 1$, according to the selected attack subset, the diffusion effect of the defense strategy set updates the weight of the attack strategy set (the stronger the diffusion effect, the greater the weight), and the cumulative weight is used to determine the random strategy at $t + 1$. After a finite number of games, the approximate solution of an optimal random defense strategy is determined by suppressing the time series of the subset.

The evaluation method that combines the model of Figure 2 and the attack graph model uses the situation value of each path as a measurement standard to determine the optimal path of the attacker and infer the attacker's attack intention. In the attack graph, each node represents the security status of the host. Therefore, each node is used as a state variable, and the vulnerability information and atomic attack behavior used in the attack process are used as observation variables.

The security event collection process collects multiple types of original event information, and the format and content of these original events are not the same. Therefore, it is necessary to format the collected security events. This paper proposes a data standard method and process based on data format and data mapping script. Data formatting scripts are used to flexibly split and assemble data as needed to achieve data formatting. The data mapping script is used to semantically express the formatted data to realize data mapping.

Table 3: Database server parameters.

Features		Service accuracy	Explanation accuracy
Database accuracy	Database Server 1	1.32	1.46
	Database Server 2	1.07	1.17
	Database Server 3	0.82	0.88
	Database Server 4	0.57	0.59
	Database Server 5	0.32	0.3
	Database Server 6	0.07	0.01
	Database Server 7	−0.18	−0.28

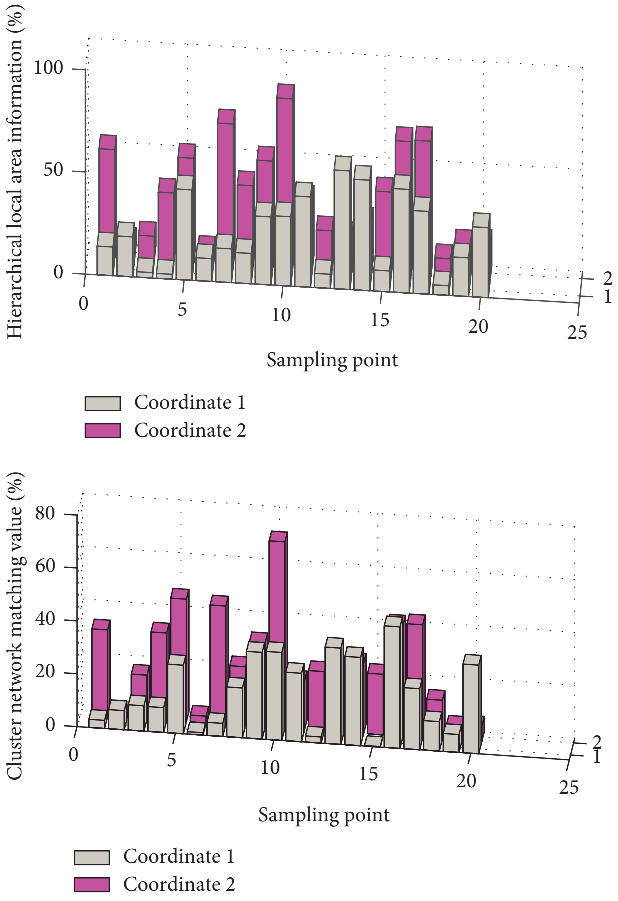

Figure 2: Hierarchical local area network matching distribution.

3.3. Security Risk Event Prediction. The foundation is a key link. The arc between the variables represents the direct causal relationship of each event; the node has a group of variables representing the state and is related to the nodes connected to it through the conditional probability matrix. The hierarchical local area network also uses the semantic hierarchical local reasoning logic, which can better reflect the easy-to-understand reasoning process, so it has also been widely used in reasoning and decision-making problems with inherent uncertainty. The lower tab bar is the primary navigation, and the upper category bar is the secondary navigation.

$$\overbrace{\left[F_i(x)+F_i(y)\right],\left[F_i(x)-F_i(y)\right]}^{x+y}=\coprod(x+y)\left[\frac{F_i(x)}{F_i(x)+F_i(y)},\frac{F_i(y)}{F_i(x)+F_i(y)}\right]. \tag{14}$$

For the filtered security incidents, there are still many duplicate or similar incidents. Therefore, the events need to be merged. The merging rules are under what circumstances, what conditions are met, and which fields are merged. The event merging function can merge a large number of security events based on merging conditions to simplify security events.

Distributed fusion is a kind of data fusion within a region. Multiple subprocessing nodes send the processed data to the sink node, and the specific information is stored locally. Finally, the sink node in Table 4 performs the data fusion. Compared with centralized fusion, it reduces the amount of communication and energy consumption, but the accuracy of fusion is lower.

It divides vulnerabilities into 26 categories, including SQL injection, code injection, trust management, information leakage, encryption issues, authorization issues, digital errors, etc.

Although other methods have fewer feature dimensions than mRMR-IG and the modeling time is relatively short, the accuracy rate is reduced and the false alarm rate is also higher than the method used in this article. It can be seen that the number of features is not as good as possible. Although fewer features reduce the modeling time and pursue time optimization, the classification information carried will also be reduced, resulting in a decrease in accuracy. Compared with all the features, although the false alarm rate has increased, with the modeling time shortened by nearly 50%, the accuracy rate is basically the same as that of all features, and the false alarm rate is reduced.

4. Big Data-Driven Hierarchical Local Area Network Security Risk Event Prediction Model Application and Analysis

4.1. Hierarchical Local Area Network Big Data Preprocessing. First of all, we obtain the data of the hierarchical local area network vulnerability. This article uses the vulnerability data in the US National Vulnerability Database NVD and the Chinese National Vulnerability Database CNNVD as the experimental data. Since the vulnerabilities in the NVD database have no category labels, the corresponding vulnerability categories in CNNVD are used.

$$\frac{\partial F_i(x)F_i(y)}{\partial F_i(x)}+\frac{\partial F_i(x)F_i(y)}{\partial F_i(y)}+\frac{\partial F_i(x)F_i(y)}{\partial (x,y)}=0. \tag{15}$$

Second, we perform word segmentation on the vulnerability description text and remove stop words to reduce data redundancy. Third, the S-C algorithm (a feature extraction algorithm based on information entropy S and comprehensive function C) is used to extract the feature word set of the vulnerability sample; finally, the word vector of each vulnerability sample is established through the feature set to complete the preprocessing of the vulnerability text data. We can see the consistency between the navigation status of the mobile phone and the navigation status of it.

The number of parameters that the particle swarm optimization algorithm needs to evolve is 42. From the real-time record of the NSFOCUS risk scan, we collected a total of 998 data. The vulnerability collection submodule is used to collect scan information of the vulnerability scanning system. After the user selects vulnerability collection, the user selects the vulnerability scan report in the xls format that needs to be imported, and after processing by the system, the imported result will be displayed on the system interface.

Considering the complexity of the large-scale network in Figure 3, it should be divided into service layer, node layer, local area network layer, and large-scale network layer. Asset assessment, threat assessment, vulnerability assessment, and risk value calculation are carried out in the hierarchy, and comprehensive risk assessment is carried out according to the relationship between the levels. Risk assessment factors include asset CIA attributes, physical value, vulnerability information, and security incidents involved in risk assessment.

The situation of network security has nothing to do with the defense capabilities of network security but only related to security incidents such as network attacks. The method of predicting the time of attacking various vulnerabilities from historical attack behaviors, using the comparison of attack time and defense time to predict whether the attack will occur or not, defines the attack time matrix and defense time matrix in the state attack graph to achieve future risks dynamic prediction. The intrusion detection system monitors the attack behavior in the network in real time, predicts the time when other atomic attacks are successfully attacked, and determines the time matrix of each state node being attacked.

$$\psi(u)=\int_0^u\left[\psi'(u)du-\psi'(0)\right]-\int_0^u \exp\left[\frac{\alpha-\beta}{b^2}\int_0^u\frac{1}{\pi(\theta)}d\theta\right]du. \tag{16}$$

Using the time when the vulnerabilities corresponding to each state node are successfully repaired to determine the defense time matrix of the state node. By comparing the attack and defense time to judge whether the attack can happen successfully and the probability of occurrence, combine the network assets to achieve the future network risk value prediction.

4.2. Realization of Network Security Risk Event Prediction Simulation. We select 700 sets of data for network training and 298 sets of data for network testing. We use the trained RBF neural network to evaluate the future risks. This shows the distribution of the difference between the estimated risk value and the actual risk value. It can be seen from the Figure 4 that the error is mainly concentrated in the 0 value. This shows that the RBF neural network has high accuracy for risk estimation. In order to test the performance of the RBF neural network, we used the BP neural network to process this set of data and compared the results. As shown in the text, we can see that the RBF neural network is more accurate.

$$\frac{\sum Y\left[i|i=0,1,2,3,,,j-1,j\right]_j}{\sum_{i=1}^{N} N\times W_j}-\frac{\sum_{i=1}^{N} X_i\times W_j}{N}=0. \tag{17}$$

It can be clearly seen from the Figure 4 that the average accuracy of using the SVDBN method is the highest. In

Table 4: Description of network node communication transmission volume.

Steps	Algorithm consumption	Description code
1	For the filtered security	Import numpy as np
2	The event merging function	For (int $i = 101$; $i < 200$; $i+ = 2$)
3	A large number of security events $\pi(u)$	Def nms (dets, thresh):
4	Accuracy of security events	Public static void main (String[] args) {
5	The accuracy of fusion is lower $\beta u + A$	Int $f1 = 1$, $f2 = 1$, f;
6	Compared with centralized fusion $\alpha - r$	Int $M = 30$;
7	Communication and energy consumption $\alpha - \beta/b$	System.out.println (f1);
8	Data-driven hierarchical local area $\alpha - \beta$	$X1 =$ dets[:, 0]
9	The merging rules are under $\psi'(u)$	$Y1 =$ dets[:, 1]
10	Connect the database through vulnerability $1/\alpha - r$	$X2 =$ dets[:, 2]
11	Environmental condition error $B * \alpha - \beta/b$	$Y2 =$ dets[:, 3]
12	It reduces the amount of $\psi''(u)$	Scores = dets[:,4]
13	A group of variables representing $(\beta u + A)^2$	For (int $i = 3$; $i < M$; i++) {
14	Many duplicate or similar incidents $\psi''(u)/\psi'(u)$	$F = f2$;
15	The false alarm rate and false alarm rate $\psi'(0)$	$F2 = f1 + f2$;
16	Other methods have fewer feature b^2	Order = scores.argsort()[-1]
17	Vulnerability category labels	Keep = []
18	System command injection $\alpha - \beta/b^2$	While order.size > 0:
19	Access verification error $\pi(\theta)$	I = order[0]
20	It divides vulnerabilities into $\int_0^u 1/\pi(\theta)$	Keep.append(i)

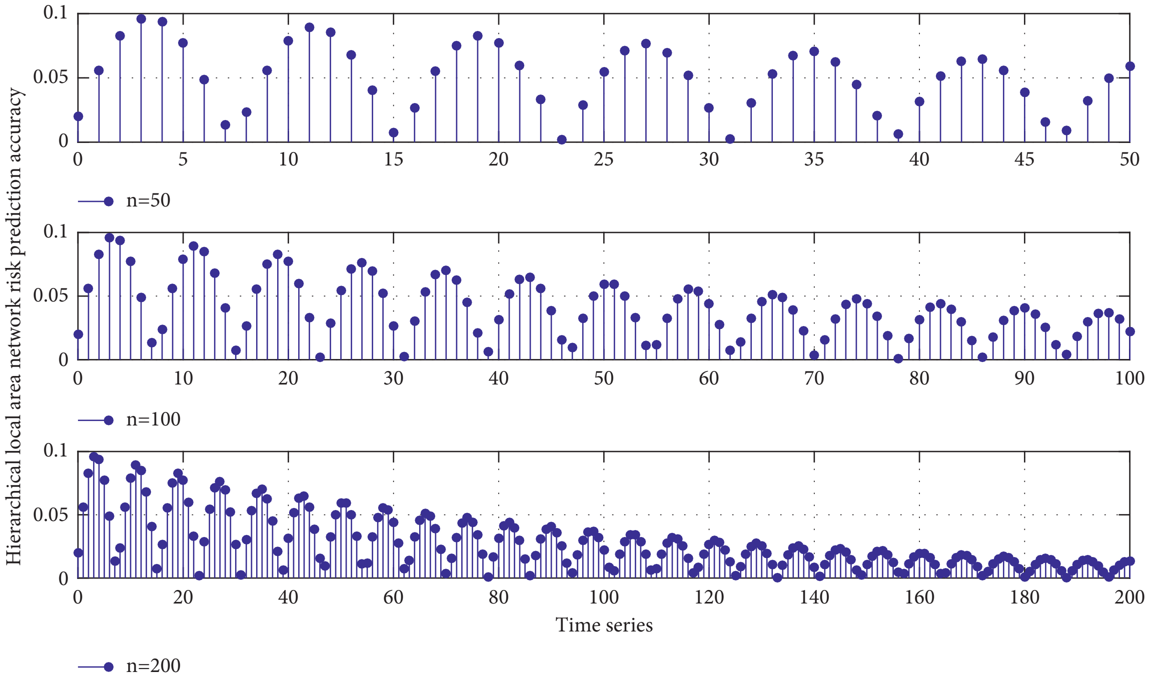

Figure 3: Distribution of hierarchical local area network risk prediction.

general, the average accuracy of using the ADBN method is slightly higher than that of the traditional BN method, but the difference is not very obvious. The reason is that the SVDBN method is better applicable to the situation of constantly changing data, where the changes between the event contexts are the changes in the free and congested traffic state. As the time interval increases, the number of unobserved data obtained increases, which will affect the performance of the model and make it difficult to achieve accurate predictions. Therefore, the average accuracy of the three methods decreases.

Selecting the feature word set that plays an important role in the vulnerability classification in Figure 4, the vector space of the vulnerability sample on the basis of this word set is built, and then the machine learning algorithm is used to effectively analyze it. Text feature extraction refers to selecting keywords that can effectively describe the category information of the vulnerability from the

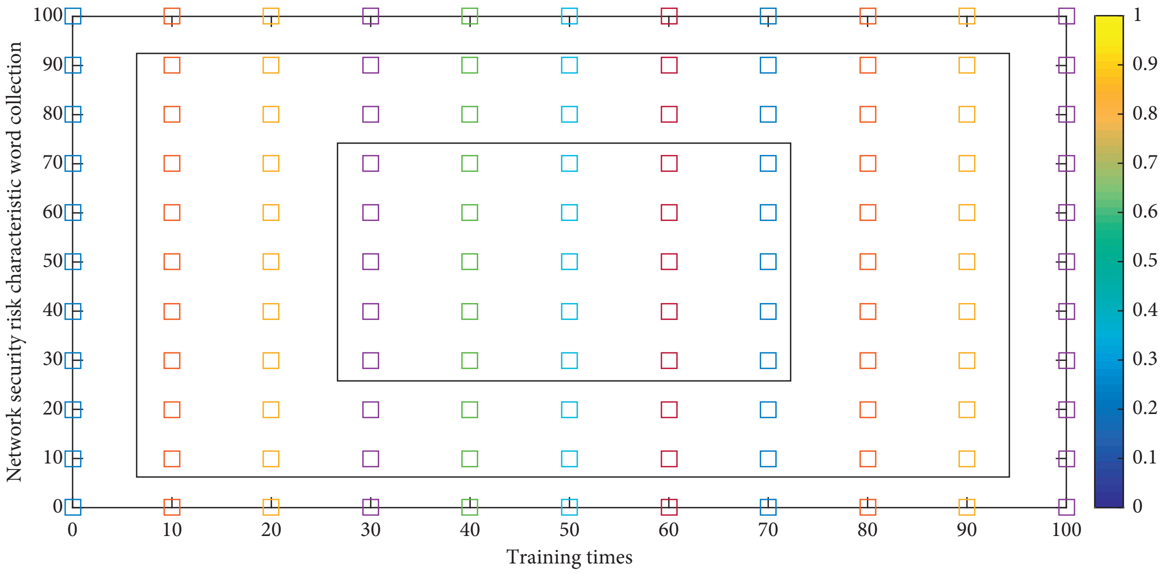

Figure 4: Network security risk characteristic word set distribution.

words described in the leaked text based on certain evaluation criteria. The huge feature word set makes the vulnerability samples generate huge-dimensional vectors during vectorization, and the efficiency and accuracy of subsequent text mining using machine learning algorithms become extremely low.

This paper analyzes the characteristics of the vulnerability-like text, defines a comprehensive function C to express the importance of words to the category, and combines the information entropy S to finally determine the S-C algorithm to extract the feature set of the data set. Although the model using the ADBN method is an adaptive model, it can also cluster historical data, learn the corresponding model from the clustered data, and select the appropriate model at runtime.

$$\lim_{i,j\longrightarrow\infty} \alpha(i,j) = \frac{\cos\left(T_i(i,j)/T_j(i,j)\right)}{\cos\left(T_i(i,j)/(i+j)\right)} \times \frac{\sin\left(T_i(i,j)/T_j(i,j)\right)}{\sin\left(T_i(i,j)/(i+j)\right)}. \tag{18}$$

We use the situation value of the attack path as a measure of risk assessment, and the probability of the attack path must be obtained before calculating the situation value. The basic principle of the algorithm is as follows: in the chain of Figure 5, there are N situations in the state at each moment, corresponding to a specified observation state. Since the state transition sequence is invisible, this must be considered in each iteration.

The actual value of each asset is the difference between the value of the asset and the value of the asset lost due to threats. The value of the asset loss is the current success of the attack. The product of the threat impact factor of the assets involved is related to the asset value of the network security incidents. From this, the actual value and loss value of the total assets in the current management domain can be calculated.

4.3. Example Application and Analysis. When judging the possibility of an attack, not only the difficulty and cost of the attack need to be considered, but also the benefit of the attack cannot be ignored. When the asset value of a subnet is greater, the attacker's intention is obviously greater, and the initial node in the state attack graph is more likely to be used by the attacker. Therefore, when the initial probability of the initial node in the state attack graph is determined, it is proportional to the size of the network's assets.

$$\begin{cases} T(i,j) = \left(\left(T_i(i,j)\right)^2 + \left(T_i(i,j)\right)^2\right)^{\frac{1}{2}}, \\ S(i,j) = \left(\left(T_i(i,j)\right)^2 - \left(T_i(i,j)\right)^2\right)^{\frac{1}{2}}. \end{cases} \tag{19}$$

And it improved the shortcomings of the independence of feature words in the naive hierarchical local classifier and correlated the vulnerability feature words to make the classifier more suitable for vulnerability samples. The vectorized vulnerability sample set in Figure 6 is used to use the AODE vulnerability classification algorithm to realize the category prediction of the missing samples in the test set.

Based on the definition of the comprehensive function C representing the importance of the feature word, the information entropy S of the word is used to weaken the importance of the more confusing words in the classification, and the word composition with a large SC value is selected as feature set. Combined with the average first-order dependency hierarchical local area algorithm,, the relationship between feature words is correlated.

In order to verify the accuracy of the vulnerability category prediction model proposed in this article, this experiment was performed on a PC with Windows 7 operating system, Intel (R) Core (TM) i7-4510U processor, clocked at 2.60 GHz, and 8.00 GB of memory. On the

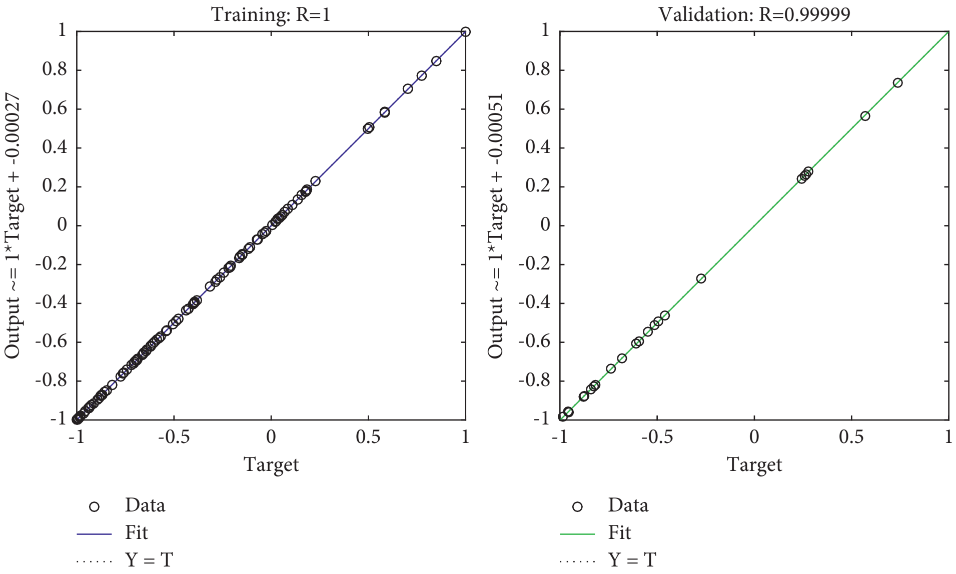

Figure 5: Distribution of hierarchical local area network state transition sequence.

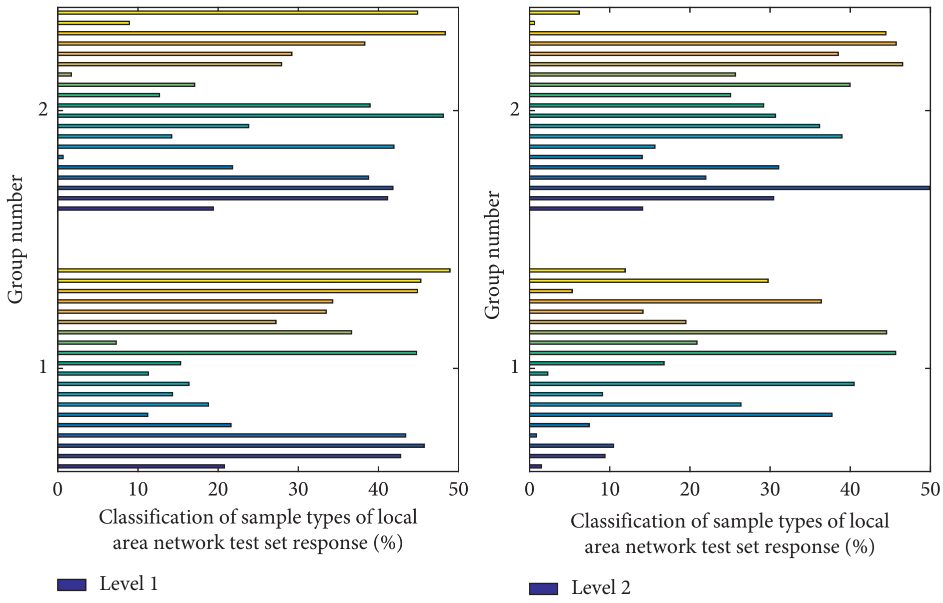

Figure 6: Prediction distribution of sample categories in the hierarchical local area network test set.

Anaconda3 version of Spyder, use python language for programming to build a vulnerability category prediction model. For a large number of unknown vulnerabilities, a vulnerability category prediction model is established. Due to the many types of vulnerabilities and the short description text, it is difficult to extract feature words. Therefore, this paper proposes the SC feature extraction method. In the design of the navigation bar, the consistency of the focus state of the navigation bar on the same page is used to facilitate user identification.

In view of the previous risk assessment methods based on state attack graphs, the previous methods ignored the impact of comprehensive factors such as attack revenue, attack capabilities, and vulnerability release time on the probability of attack events and also ignored risk prediction. The model uses comprehensive factors to determine the status transition matrix, making risk quantification more comprehensive and reasonable. The reachability probability of state nodes is dynamically updated in real time by using captured intrusion behaviors, and the dynamic risk assessment and prediction are realized by determining the attack time matrix and defense time matrix.

Each state node in the state attack diagram in Figure 7 represents the authority of each host that an attacker can obtain by exploiting the vulnerability. The risk value of the network is obtained by adding the risk value of the host, and the risk value of the host is obtained by adding the risk value caused by the vulnerability of the state node related to it. To

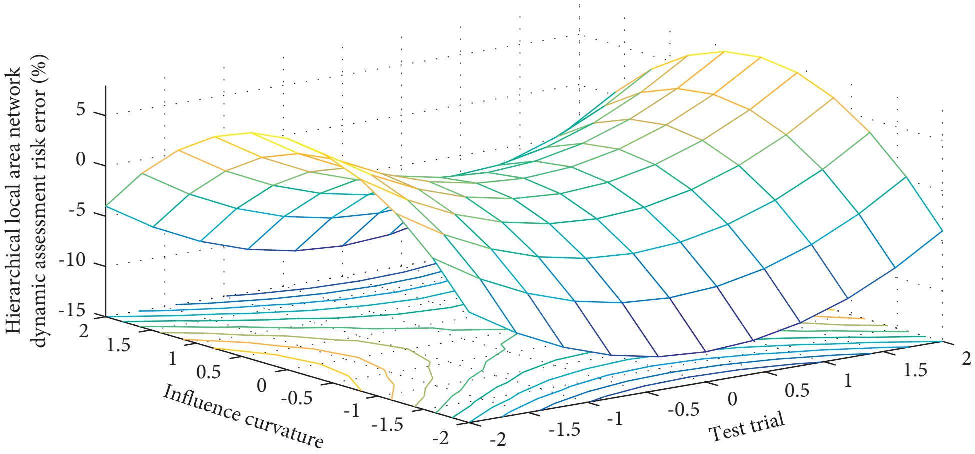

Figure 7: Dynamic assessment of risk distribution in hierarchical local area network.

quantify the risk value of the state node, it is necessary to calculate the probability that the attacker can successfully attack to reach each state node and finally use the asset damage value caused by the relevant vulnerability to multiply the asset value of the host node to quantify the risk.

$$\iint \sin\,\theta\,\cos\,\theta R(Gau(ss))dsd\theta = \iint \sin\,u\,\cos\,u\,\sin\,\theta\,\cos\,\theta\,dud\theta. \tag{20}$$

This article uses the metrics and values of the indicators specified in the vulnerability scoring framework CVSS of the US National Vulnerability Database NVD as a reference to define how easy it is to exploit the vulnerability. This model uses factors such as attack revenue, attacker capabilities, vulnerability release time, and other factors to jointly determine the state transition matrix, so that the reachability probability of state nodes is calculated more accurately.

The attack behavior detected by the intrusion detection system is used to determine the attacker's ability, attack time matrix, and defense time matrix, so as to determine the attack behavior and defense behavior occurring at each moment, and dynamically update the initial state vector and state transition matrix in the state attack graph so as to realize the calculation of future risks. Through dynamic network risk assessment, the current and future risk status of the network system can be visualized.

5. Conclusion

Aiming at the problem that the hierarchical local evolution model may not be able to predict in a short time under the data-driven situation, this paper proposes a predictive complex event processing method based on a variable structure dynamic hierarchical local area network. In this method, historical data are divided by offline context clustering, and then different clusters are obtained. The data divided into each cluster use a scoring search method to compare the corresponding data. The hierarchical local area network is used for learning, and the Gaussian mixture model is used for approximate inference. First of all, based on the exchange server being applied, combined with the hardware control technology of the security isolation gatekeeper, a complete mail management system is established to realize mail between internal and external networks. Second, we apply the access control of the internal network machine and the hardware encryption function of the U disk provided by the IT security operation and maintenance system to realize the addition of client devices and file copy restrictions. Compared with the classification under all features, when only the mRMR method is used, the detection rates of Probe and R2L have been improved, but the detection rate of U2R has not increased but decreased, which shows that the redundant features removed by mRMR contain features that are highly related to U2R; when only the IG method is used, the detection rate of the Probe class is significantly improved, and the detection rate of U2R decreases relatively less, which indicates that the features screened by IG are similar to mRMR. There are more features than those related to these two categories. It can be seen that, among the redundant features removed after screening using a feature selection algorithm alone, there are features that are highly related to certain categories, and the combination of the two can largely compensate for the shortcomings of each screening method. For online data, select the appropriate hierarchical local area network model or combination of models for prediction and inference according to the current event context and dynamically update the data in the cluster when the event stream is predicted online, so that the prediction model can be updated in real time.

Data Availability

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

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the College of Modern industry (szxy2021cyxy04); School-Level Scientific Research Platform of Suzhou University (2021xjpt51); and 2019 Teaching and Research Project of Suzhou University (szxy2019jyxm12);

References

[1] T. S. J. Darwish and K. Abu Bakar, "Fog based intelligent transportation big data analytics in the internet of vehicles environment: motivations, architecture, challenges, and critical issues," *IEEE Access*, vol. 6, pp. 15679–15701, 2018.

[2] L. D. Xu and L. Duan, "Big data for cyber physical systems in industry 4.0: a survey," *Enterprise Information Systems*, vol. 13, no. 2, pp. 148–169, 2019.

[3] M. I. Razzak, M. Imran, and G. Xu, "Big data analytics for preventive medicine," *Neural Computing and Applications*, vol. 32, no. 9, pp. 4417–4451, 2020.

[4] Y. Zhang, T. Huang, and E. F. Bompard, "Big data analytics in smart grids: a review," *Energy informatics*, vol. 1, no. 1, pp. 14–24, 2018.

[5] H. N. Dai, H. Wang, G. Xu et al., "Big data analytics for manufacturing internet of things: opportunities, challenges and enabling technologies," *Enterprise Information Systems*, vol. 14, no. 9-10, pp. 1279–1303, 2020.

[6] N. Sun, J. Zhang, P. Rimba et al., "Data-driven cybersecurity incident prediction: A survey," *IEEE communications surveys & tutorials*, vol. 21, no. 2, pp. 1744–1772, 2018.

[7] E. Hossain, I. Khan, F. Un-Noor, S. S. Sikander, and M. S. H. Sunny, "Application of big data and machine learning in smart grid, and associated security concerns: A review," *Ieee Access*, vol. 7, pp. 13960–13988, 2019.

[8] S. Jiang, M. Lian, C. Lu et al., "Ensemble prediction algorithm of anomaly monitoring based on big data analysis platform of open-pit mine slope," *Complexity*, vol. 2, p. 18, 2018.

[9] M. Fathi, M. Haghi Kashani, S. M. Jameii et al., "Big data analytics in weather forecasting: A systematic review," *Archives of Computational Methods in Engineering*, vol. 12, pp. 17–29, 2021.

[10] L. Greco, G. Percannella, P. Ritrovato, F. Tortorella, and M. Vento, "Trends in IoT based solutions for health care: Moving AI to the edge," *Pattern Recognition Letters*, vol. 135, pp. 346–353, 2020.

[11] R. Dautov, S. Distefano, and R. Buyya, "Hierarchical data fusion for smart healthcare," *Journal of Big Data*, vol. 6, pp. 10–23, 2019.

[12] S. Bazzaz Abkenar, M. Haghi Kashani, E. Mahdipour, and S. M. Jameii, "Big data analytics meets social media: A systematic review of techniques, open issues, and future directions," *Telematics and Informatics*, vol. 57, p. 101517, 2021.

[13] Y. Yu, M. Li, L. Liu, Y. Li, and J. Wang, "Clinical big data and deep learning: Applications, challenges, and future outlooks," *Big Data Mining and Analytics*, vol. 2, no. 4, pp. 288–305, 2019.

[14] C. Tu, X. He, Z. Shuai, and F. Jiang, "Big data issues in smart grid - A review," *Renewable and Sustainable Energy Reviews*, vol. 79, pp. 1099–1107, 2017.

[15] J. Lee, J. Kim, I. Kim, and K. Han, "Cyber threat detection based on artificial neural networks using event profiles," *IEEE Access*, vol. 7, pp. 165607–165626, 2019.

[16] M. Ghorbanian, S. H. Dolatabadi, and P. Siano, "Big data issues in smart grids: A survey," *IEEE Systems Journal*, vol. 13, no. 4, pp. 4158–4168, 2019.

[17] I. Lana, J. Del Ser, M. Velez, and E. I. Vlahogianni, "Road traffic forecasting: recent advances and new challenges," *IEEE Intelligent Transportation Systems Magazine*, vol. 10, no. 2, pp. 93–109, 2018.

[18] M. Feng, J. Zheng, J. Ren et al., "Big data analytics and mining for effective visualization and trends forecasting of crime data," *IEEE Access*, vol. 7, pp. 106111–106123, 2019.

[19] T. Kolajo, O. Daramola, and A. Adebiyi, "Big data stream analysis: a systematic literature review," *Journal of Big Data*, vol. 6, no. 1, pp. 18–30, 2019.

[20] S. Brayne and A. Christin, "Technologies of crime prediction: the reception of algorithms in policing and criminal courts," *Social Problems*, vol. 68, no. 3, pp. 608–624, 2021.

[21] J. Hegde and B. Rokseth, "Applications of machine learning methods for engineering risk assessment - A review," *Safety Science*, vol. 122, p. 104492, 2020.

[22] B. P. Bhattarai, S. Paudyal, Y. Luo et al., "Big data analytics in smart grids: state-of-the-art, challenges, opportunities, and future directions," *IET Smart Grid*, vol. 2, no. 2, pp. 141–154, 2019.

[23] S. B. Atitallah, M. Driss, W. Boulila, and H. B. Ghézala, "Leveraging Deep Learning and IoT big data analytics to support the smart cities development: review and future directions," *Computer Science Review*, vol. 38, p. 100303, 2020.

[24] R. Iqbal, F. Doctor, B. More, S. Mahmud, and U. Yousuf, "Big data analytics and computational intelligence for cyber-physical systems: recent trends and state of the art applications," *Future Generation Computer Systems*, vol. 105, pp. 766–778, 2020.

[25] V. Jagadeeswari, V. Subramaniyaswamy, R. Logesh, and V. Vijayakumar, "A study on medical Internet of Things and Big Data in personalized healthcare system," *Health Information Science and Systems*, vol. 6, no. 1, pp. 14–20, 2018.

[26] A. Y. Sun and B. R. Scanlon, "How can Big Data and machine learning benefit environment and water management: a survey of methods, applications, and future directions," *Environmental Research Letters*, vol. 14, no. 7, p. 073001, 2019.