Applied Mathematics, Modeling and Computer Simulation
C. Chen (Ed.)

doi:10.3233/ATDE220097

Based on Computer Immune Algorithm Model Big Data Security Audit Research

Xin Liu[a], Bin Luo[a,1], Zhigang Guo[a], Wenfeng Xu[a] and Xiaodong Yu[a]
[a] *Hubei Central China Technology Development of Electric Power Co.Ltd 430077*

Abstract. In recent years, with the rapid development of internet big data technology and applications, data security issues have become increasingly serious, and incidents of infringement and theft of citizens' privacy and sensitive data have emerged endlessly. This article mainly uses the computer immune algorithm model to conduct some in-depth research on the big data security audit problem. Data security refers to the use of multiple operation and maintenance technologies to ensure the normal operation of the data system, thereby ensuring the availability, integrity, and confidentiality of data. In the era of the rapid development of the internet, data security is a problem faces by all major companies, especially some large companies such as power companies, which have huge corporate databases and complex data volumes. Once a data security incident occurs during the operation of an enterprise, it will have a major impact on the business and normal operation of the enterprise. At present, in order to effectively avoid data security issues, many companies have adopted security inspections, graded protection, security audits and other methods to deal with data security issues. At present, data security auditing has become more and more popular among enterprises of a certain size, and related work has been started. Data security audits have gradually become the primary technical means for enterprises to prevent data security issues.

Keywords. Big data technology, big data security audit, immune algorithm model.

1. Introduction

With the rapid development of data informatization, public privacy data is frequently stolen and leaked major data security accidents. Big data is usually stored in a distributed way due to the huge amount of stored data. In this way, the path view of storage is relatively clear, and data protection is too simple due to the large amount of data. Hackers and other factors may easily take advantage of relevant vulnerabilities and cause security problems. In the big data environment, there are many end users and many types of users, which requires a lot of manpower and time for user identity authentication[1]. Once the data is attacked, the output data of the big data platform may be intercepted, resulting in great information security risks.

To solve this problem, we can learn from the working mode of the human immune system to achieve data security protection. Biological immune system is an adaptive complex system with distributed, self-organizing and dynamic balancing capabilities. The immune system invades the antigen to the outside world, can produce the corresponding antibody by the lymphatic cell of different kinds that distributes the

[1] Corresponding Author, Bin Luo, Hubei Central China Technology Develoment of Electric Power Co.Ltd, China; E-mail: 1540567431@qq.com.

whole body. In order to ensure the normal operation of the basic physiological functions and activities of the whole biological system. It has a good ability of classification detection, especially for the analysis and treatment of multi-modal problems with high intelligence and robustness.

This paper mainly describes the relevant concepts and mechanisms of big data security audit, related concepts and mechanisms of immune algorithm and its application mechanism in big data security audit, and discusses the improvement and optimization of immune algorithm in the application
of big data security audit.

2. Data Security Issues

2.1. Data Security

On June 10, 2021, the 29th Meeting of the Standing Committee of the 13th National People's Congress adopted the Data Security Law of the People's Republic of China, which will come into force on September 1, 2021. The Data Security Law contributes Chinese wisdom and solutions to global data security governance.

Data security refers to the necessary means to ensure that the data is in the state of effective protection and legal use, and has the ability to guarantee the continuous security state.

2.2. Data Security Threat Factors

There are many factors that threaten data security. The common ones are as follows:

- Hacker: intruders remotely invade the database system through the network with the help of system vulnerabilities.
- virus: the computer due to the infection of the virus caused damage, computer virus is highly infectious, especially in the network environment, the spread of faster. As a result, related data is lost and leaked.
- Steal information: copy, delete information from the computer or steal the computer.
- Human error: due to operational errors, users may mistakenly delete important files of the system, or fail to operate in accordance with relevant requirements and crash the data system.
- Hard drive damage: If the hard drive is physically damaged, data may be lost. Disk drives are affected by factors such as device running loss, storage media failure, and human sabotage.

2.3. Big Data Security Audit

2.3.1. Concepts Related to Big Data Security Audit

The threats and risks faced by data are dynamic processes. The intrusions, methods and targets change over time. This calls for the continuous strengthening of our protection system, which cannot remain static and must be able to effectively respond to this problem[2]. So in the process of data security protection to have a crucial ability, is perfect data audit ability. The audit capability helps us understand the changes of

threats and risks, clarify the direction of protection, further adjust the protection system, optimize the defense strategy, strengthen the weak points in the defense, and make the defense system have dynamic adaptability, so as to achieve the real data security protection capability.

2.3.2. Big Data Security Audit

Big data security audit consists of behavior audit and analysis, permission change monitoring and abnormal behavior analysis.

Behavior audit and analysis refers to the use of database protocol analysis technology to record all the behavior of accessing and using data. Perfect data security audit system can bring us alarm in the event and trace the source of the two mechanisms. In case of malicious behaviors that may lead to data leakage during data access and use, the audit system sends a threat alarm to notify the staff immediately. The staff can deal with the threat in time, so as to minimize or even avoid the loss. In order to realize the alarm function, audit system is required to identify risks and threats effectively. After a security event occurs in the process of data access, the audit system can trace the event to the source, which is to determine what causes the event and when and where it happens. A preliminary reduction is made to the approximate occurrence process of the event and the cause of the event is analyzed. It provides valuable reference for adjusting and improving data protection measures. Therefore, audit system must have good retrieval ability, so as to locate and trace the source after the event. Audit system is based on abnormal data, data pipeline and other aspects of the assembly, the data flow, data operation monitoring, audit analysis, timely discovery of abnormal data and data operation and alarm[3].

Permission change monitoring is to monitor the changes of all data access permissions. By monitoring permission changes, it can timely resist malicious attacks from the outside, and also detect and block related permission adjustment violations from the inside. There are two stages in the establishment of the monitoring of authority change, one is authority combing, the other is authority monitoring[4].

Abnormal behavior analysis refers to the study and modeling of the data in the daily audit system, and then the data in the real-time monitoring system is compared and analyzed with the daily data. If the monitored data exceeds a certain error range, the feedback will be given to the staff, so as to make early warning and data protection measures in advance. Figure 1 shows the big data security audit process[5].

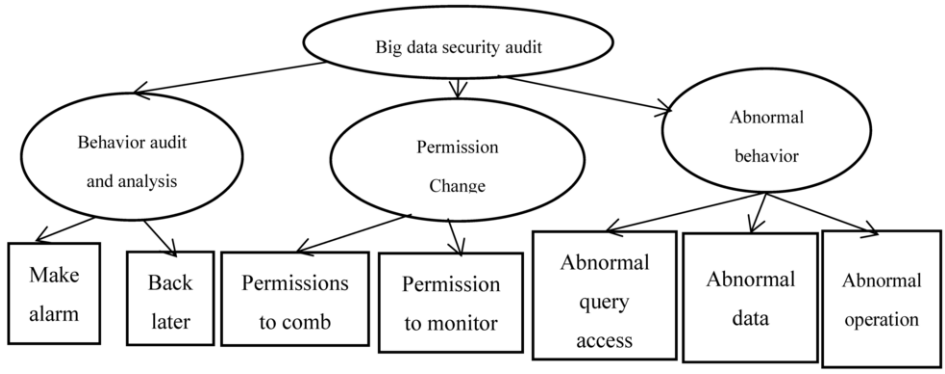

Figure 1. Big data security audit procedure

3. Computer Immune Algorithm Model

3.1. Principles of Biological Immunology

Immunity is a physiological function of the human body, according to which the human body can achieve the recognition of its "own" and "non" components. Further damage prevents antigens such as viruses from entering the body or the bad cells and tumor cells produced by the body to maintain health. Generally speaking, immunity refers to a physiological protection function of the organism. It includes a series of processes such as the screening, elimination and elimination of foreign bodies. To sum up, the function of immune system is mainly manifested in three aspects, namely defense function, stability function and immune monitoring function. Figure 2 shows the general process of immunity[6].

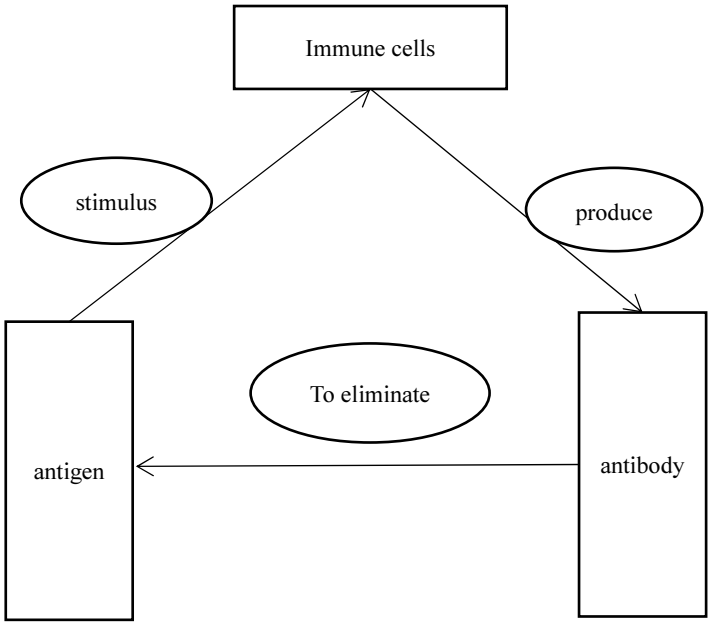

Figure 2. General process of immunity

3.2. Development History of Immune Algorithms

The working mechanism of human immune system attracts many researchers to build algorithm model to solve some problems in real life. In the field of life sciences, natural phenomena such as genetics and immunity have been studied in some depth. In the 1960s, Bagley and Rosenberg and other scientists successfully applied their knowledge, especially some theories of genetics, to some fields of engineering science on the basis of analyzing and understanding the existing research results. Professor Hollan summarized and generalized the concept of heredity, and gave a concise description of genetic algorithm. It laid the theoretical foundation for the later immune algorithm. In 1958, Australian scholar Burnet first proposed the clonal selection theory related to

immune algorithm, which laid a solid foundation for the subsequent research on immune algorithm. In 1973, Jerne proposed the immune system model, created a unique network theory based on the principle of clonal selection, gave the mathematical framework of the immune system, and used differential equation modeling to simulate the dynamic changes of lymphocytes. In 1986, Farmal et al. constructed a dynamic model of the immune system based on the immune network theory, indicating the possibility of the immune system being combined with other artificial intelligence methods, and initiating the study of immune algorithm. Immune Algorithm is a swarm intelligence search Algorithm with an iterative process of generate and test[7].

3.3. Immune Algorithm Model

In the immune algorithm, antigen can be simply understood as a problem to be solved, antibody as a specific solution vector, antigen recognition as problem recognition, lymphocyte differentiation as the preservation of excellent solutions, cell inhibition as the elimination of remaining candidate solutions, and cell cloning as the generation of new antibodies by genetic operators[8].

3.3.1. Immune Algorithm Module

1. Antigen recognition and initial antibody production. According to the characteristics of the problem to be optimized, appropriate antibody coding rules are designed and the priori of the problem is used to generate the initial antibody population.

2. Antibody evaluation: The quality of antibodies is evaluated. The evaluation criteria mainly include antibody affinity and antibody concentration.

3. Immune operation.

Immune operator plays an important role in immune operation. The affinity operator represents the degree of binding between immune cells and antigen (that is, the degree of binding between feasible solution and the problem to be solved). Usually function optimization problems can use the value of the function or the reciprocal of the function value as the evaluation of affinity. The affinity operator is usually expressed by the function AFF (x). Antibody concentration is a characterization of the diversity of antibody populations. The high concentration of antibody means that there are a large number of very similar individuals in the population, and the optimization will be concentrated in a region of a feasible interval, which is not conducive to global optimization. Antibody excitation is the final evaluation result of antibody quality, which requires comprehensive consideration of antibody affinity and antibody concentration. Generally, antibodies with high affinity and low concentration will get higher excitation degree. The immune selection operator selects those antibodies according to their excitation to enter the clonal selection operation. Antibody individuals with high excitation have better quality and are more likely to be selected for clonal selection[9].

3.3.2. Immune Algorithm Process

1. Carry out antigen identification, that is, to understand the optimization problem —— antigen corresponds to the problem to be solved. The problem is analyzed, the prior knowledge is proposed, the appropriate affinity function is constructed, and various constraints are restricted.

2. Generate the initial antibody group, and express the feasible solution of the problem as the antibody in the solution space through coding, and randomly generate an initial population in the solution space (in fact, the solution space is the value of X, and we set the initial population by finding several starting positions in the solution).

3. Evaluate the affinity degree of each feasible solution in the population.

4. Determine whether the termination conditions are met, terminate the algorithm optimization process, and output the calculation results.

5. Calculate the antibody concentration and excitation degree.

6. Immunize.

Immune selection: The selection of high quality antibodies based on the affinity and concentration of antibodies in the population for activation.

Clone: clone the activated antibody and get several copies.

Mutation: to clone the copy of the mutation operation, so that its affinity mutation.

Clonal inhibition: reselection of mutation results, inhibition of low affinity, retention of high affinity antibodies.

7. Population refresh

Replace the low excitation antibodies in the population with randomly generated new antibodies to form a new generation of antibodies, go to Step 3.

Figure 3 shows the flow chart of immune algorithm.

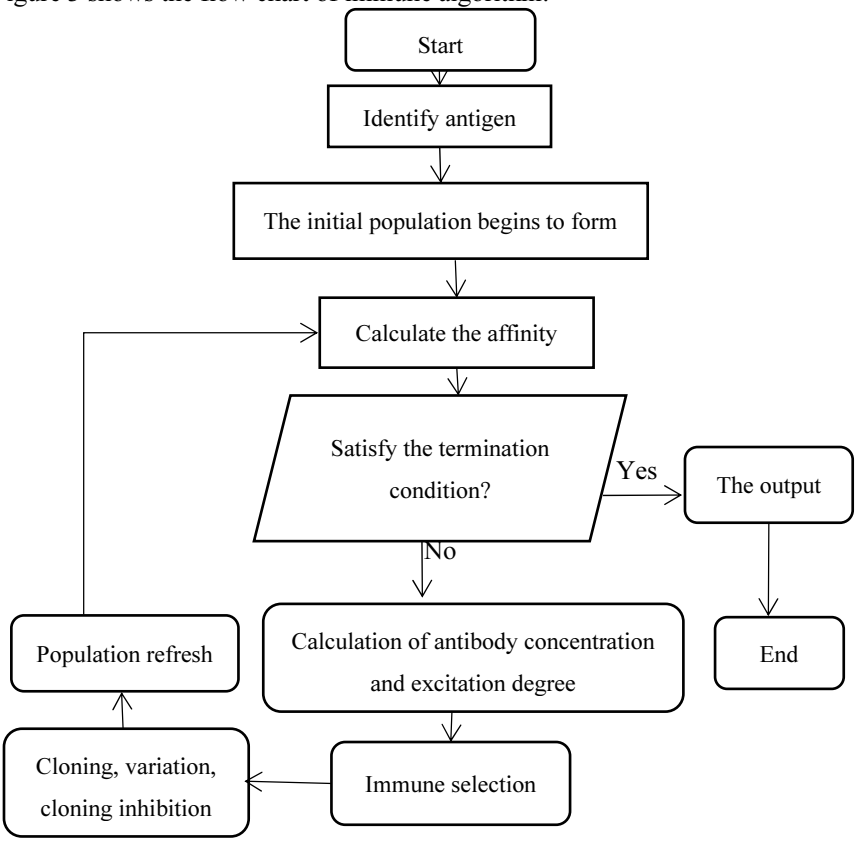

Figure 3. Flowchart of immune algorithm

3.3.3. Definitions In Immune Operations

Definition of antibody concentration:$den(ab_i)=\frac{1}{N}\sum_{j=1}^{N} S(ab_i,ab_j)$. Where N is the population size, $S(ab_i,ab_j)$ is the similarity between antibodies, which can be expressed as:

$$S(ab_i,ab_j)=\begin{cases}1, aff(ab_i,ab_j)<\delta s\\ 0, aff(ab_i,ab_j)\geq\delta s\end{cases} \quad (1)$$

Where ab_i is the ith antibody of the population, $aff(ab_i,ab_j)$ is the affinity between antibody i and antibody j, and δs is the similarity threshold.

There are two methods to calculate the affinity between antibodies (representing the degree of similarity between antibodies) : one is the method of affinity between antibodies based on Euclidean moment separation; the other is the method of antibody-antibody affinity calculation based on Haiming moment separation.

Calculation method of affinity between antibodies based on Euclidean distance:

$$aff(ab_i,ab_j)=\sqrt{\sum_{k=1}^{L}(abi,k-abj,k)^2} \quad (2)$$

Where, are the Kth dimension of antibody i and the Kth dimension of antibody j respectively, and L is the total dimension of antibody encoding.

Calculation method of affinity between antibodies based on Hemming distance:

$aff(ab_i,ab_j)=\sum_{k=1}^{L}\partial_k$,Among them:

$$\partial_k=\begin{cases}1, ab_{i,k}=ab_{j,k}\\ 0, ab_{i,k}\neq ab_{j,k}\end{cases} \quad (3)$$

$ab_{i,k}$ and $ab_{j,k}$ represent the k position of antibody i and the k position of antibody j respectively, and L is the encoding length of antibody.

The Hemming distance is 1 if the two antibody attribute values are the same, otherwise it is 0, and then the sum operation is carried out. It's just the number of numbers that have exactly the same internal properties.

The expression of antibody excitation is: $sim(ab_i)=a\cdot aff(ab_i)-b\cdot den(ab_i)$, Where $sim(ab_i)$ is the excitation degree of antibody ab_i; a and b are calculation parameters and can be determined according to the actual situation.

The clonal operator replicates the antibody selected by the immune selection operator, which can be expressed as:

$T_c=(ab_i)=clone(ab_i)$, Where $clone(ab_i)$ is the assembly of m_i clones identical with ab_i, m_i is the number of antibody clones, which can be determined previously or dynamically and adaptively calculated.

The mutation operator can mutate the result of antibody cloning obtained by the clone operator, so as to generate affinity mutation and realize local search, which has a great influence on the performance of the algorithm. Real coding algorithm and discrete coding algorithm adopt different mutation operators.

Real coding algorithm is to add a small disturbance to the source of variation to realize the source of variation neighborhood search. Can be expressed as:

$$T_m(ab_{i,j,m})=\begin{cases} ab_{i,j,m}+(rand-0.5)\cdot\delta, & rand<p_m \\ ab_{i,j,m}, & other \end{cases} \quad (4)$$

Where, $ab_{i,j,m}$ is the j-th dimension of the m-th clone of antibody abi, δ is the neighborhood range, which can be determined in advance or adjusted according to the actual evolution, rand is the random number function within the range of (0,1), and p_m is the mutation probability.

The discrete coding algorithm mainly uses binary coding, randomly selects several bits from the mutant source antibody, inverts the selected element value, and makes it in the mutant source neighborhood in the discrete space.

4. Application Mechanismof Immune Algorithm In Big Data Security Audit

4.1. Change Perception Method

4.1.1. Introduction to Hazard Theory

Danger theory states that the immune system does not simply identify self and non-self, but harmful self and non-self, and that what matters to the immune system is the "danger" of an invasion rather than the "foreignness" of the invasion. The immune system defends against danger signals, and only danger signals can stimulate the activation of effector cells to trigger an immune response. The danger theory holds that the immune system defuses potential dangers and recognizes "dangerous antigens."

According to hazard theory, an invading agent causes abnormal death, rather than normal apoptosis, of lymphocytes that release substances not present in tissue, called damps (hazard associated molecular patterns), which are recognized by toll-like receptors (TLRS) on antigen-presenting cells (APC). The APC activated by DAMP sends a costimulatory signal to T cells to activate T cells, which in turn activate B cells. Complete the whole process of immune cell recognition of antigen. Figure 4 shows the working principle of the hazard theory.

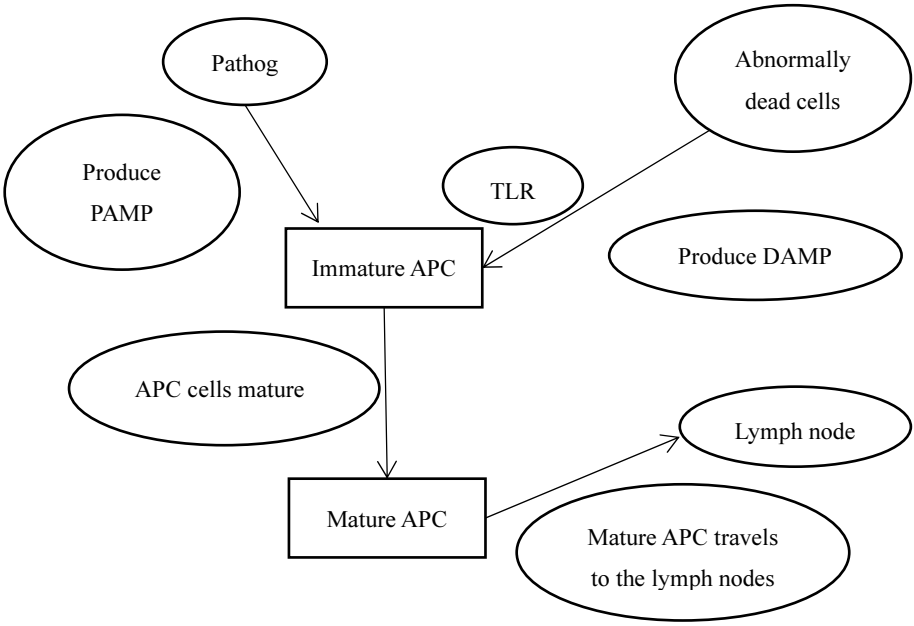

Figure 4. Schematic diagram of hazard theory

4.1.2. Principle of Change Perception

According to the danger theory, the real cause of system abnormality or collapse is the potential danger that poses a threat to the system. The immune system detects the changes in the system caused by invading antigens, detects whether the changes pose a threat to the system, and then takes necessary measures to suppress harmful changes and ignore harmless changes. When data system damage, such as hacker or Trojan virus attack, data system balance is broken, all sorts of change measure of system safety index, predicted the arrival of danger and so on based on the application of immune algorithm in big data security audit research, important is to find lead to abnormal changes of system namely "danger", Then complete the immune algorithm to the data system adaptive adjustment[10].

4.1.3. Danger Signal

Red flags should show up early in the infection, which minimizes damage and can be quickly detected. Possible red flags in a data system include the following:

- Data input and output are abnormal (for example, the transfer rate is suddenly too high or too low, and file data cannot be read or written).
- Data is abnormally lost.
- The data system was attacked and crashed, leading to failure to log in.
- Read more and write more, read less and write less, or file garbled when reading specified data.

Like individual organisms, data systems have diversity. For example, data systems are based on different operating environments and network environments, and different users make different data systems have different judgments on danger signals.

4.2. Application Model of Immune Algorithm in Data Security Audit

4.2.1. Model Structure

The model consists of data acquisition module, danger signal generation module, co-stimulus signal generation module and warning module. Figure 5 shows the data acquisition module, figure 6 shows the danger signal generation module, and figure 7 shows the relationship between modules.

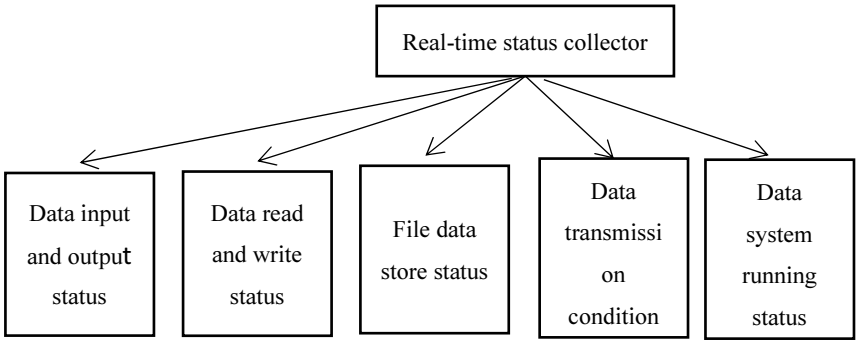

Figure 5. Data collection module

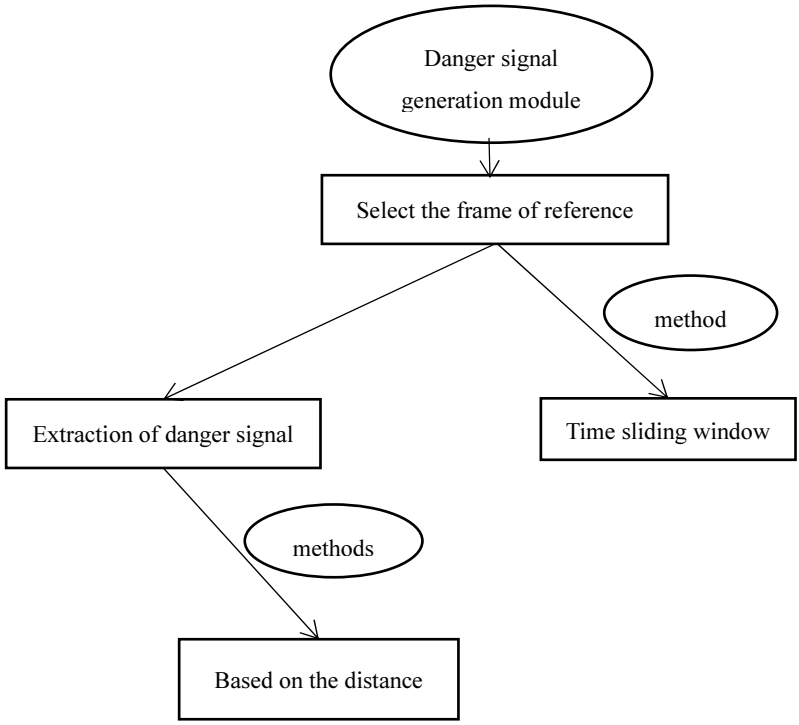

Figure 6. Module for generating danger signals

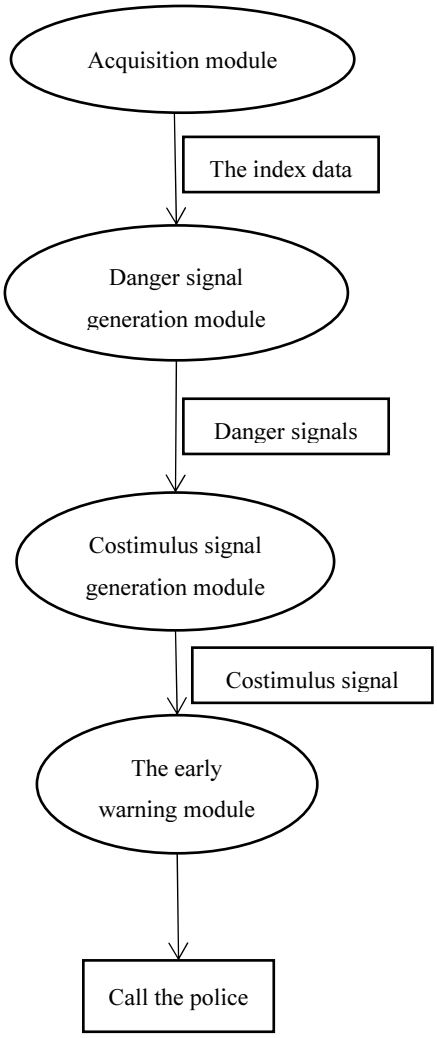

Figure 7. Relationship between modules

4.2.2. Working Mode Principle and Mechanism

In the practical application of the immune algorithm proposed by using the bionic idea for reference to the principle of human immunity, the most important thing is to make use of the characteristics of pattern recognition. For example, the recognition between antigen and antibody, immune system to "self" and "not self" recognition and so on. Based on immune algorithm thought, through the acquisition module, data flow and operation of the real-time monitoring data systems, reference behavior monitoring method principle, the use of "what is a normal" system to the detection data of abnormal behavior, i.e., data system now didn't appear abnormal problem, as long as it

has abnormal behavior, will be marked.

In the data system, there are some behaviors such as users reading and writing data, copying or deleting data, etc., which can be tracked and recorded. Data system hidden security risks most operations are similar to the above behaviors, and compared with normal operations, behavior monitoring method can use these behaviors to monitor because of their special behaviors.

Behavior monitoring method is divided into learning stage and detection stage. In the process of learning stage, we learn the normal behavior of data system operation, and build a behavior model of data system under normal conditions by learning the historical data operation behavior and data flow. In the process of the test phase, the use of the existing learned knowledge to determine whether malicious behavior, if it is found that some data manipulation behavior and to the predefined normal behavior in the variation within a certain error range, can be concluded that the abnormal behavior of fishy, and terminate the operation behavior and phase correlation operation behavior. The operation status of this behavior is reported to the maintenance personnel of the data system for further processing. Figure 8 shows the working mode schematic diagram.

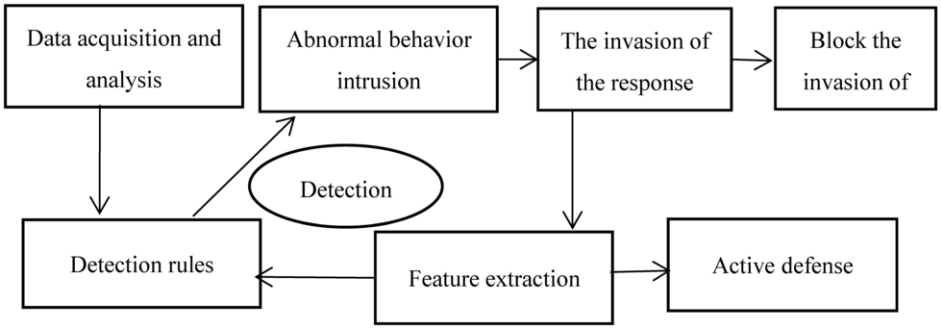

Figure 8. Working mode schematic diagram

5. Optimization of the Application of Immune Algorithm in Big Data Security Audit

5.1. Application Optimization of Immune Algorithm

Immune algorithm is a kind of imitated intelligent optimization algorithm based on biological immune mechanism, which can efficiently search for global optimal solution through the recognition effect of antibody on antigen. Above this paper expounds the application of immune algorithm in the data security audit principle and mechanism of data system once appear problems in the actual problem through the mechanism of immune algorithm model can better detect abnormal problem, but the process from discovery to solve problems of need time may be longer, and if large data system doesn't work for a long time, It will cause great impact and economic loss to individuals and enterprises, so it is necessary to optimize the immune algorithm to improve the efficiency and accuracy of detecting and solving abnormal problems.

In order to improve the optimization immune algorithm, particle swarm optimization parallel computing can be used to update the memory library, which

increases the information transmission ability of memory cells, enables better antibodies to realize message sharing, so as to avoid falling into local optimal, and provides a feasible method for improving detection efficiency and accuracy.

5.2. Particle swarm optimization

The core idea of particle swarm optimization (PSO) algorithm is to make use of the information sharing of individuals in a group to make the whole group movement in the process of solving spatial problems from disorder to order, and obtain the optimal solution of the problem. Particle swarm optimization (PSO) is a swarm cooperative random search algorithm developed by simulating bird foraging. PSO is initialized to a group of particles and then the optimal solution is obtained by iterative methods.

The overall framework of particle swarm optimization is as follows:

1. Population initialization: random initialization or specific initialization method can be designed according to the optimized problem, and then the individual adaptive value can be calculated to select the local optimal position vector of the individual and the global optimal position vector of the population.
2. Iteration setting: Set the iteration number and set the current iteration number to 1.
3. Speed update: Update the speed vector for each individual.
4. Position update: Update the position vector of each individual.
5. Local position and global position vector update: update the local optimal solution of each individual and the global optimal solution of the population.
6. Judgment of termination conditions: when judging the number of iterations, the maximum number of iterations is reached. If so, output the global optimal solution; otherwise, continue the iteration and skip to Step 3.

Figure 9 shows the flow chart of particle swarm optimization.

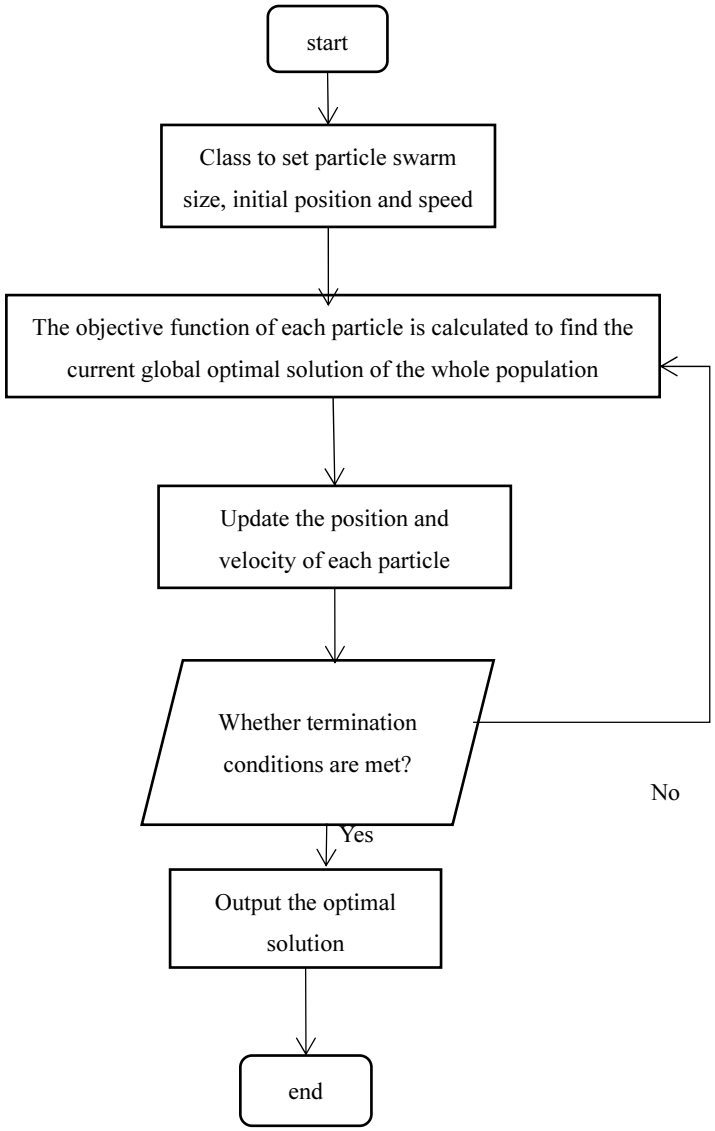

Figure 9. Flow chart of particle swarm optimization

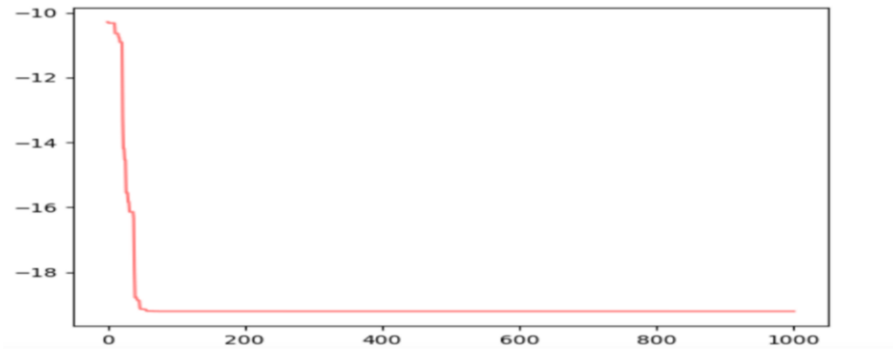

Figure 10. Experimental diagram of particle swarm optimization

The experimental figure of particle swarm optimization is shown in figure 10.

5.3. Realization of Immune Algorithm Optimization in Application

The optimized immune algorithm in this paper is based on the parallel operation method of particle swarm algorithm, and the initialization of the same two populations are divided into particle swarm algorithm and immune algorithm. The information update of the particle swarm depends only on its own operation mode, while the information update of the immune population depends on the traditional immune operation and the information differentiation of memory cells. The memory cell update of the memory bank is determined by the two algorithms together, and the specific flow chart is shown in figure 11.

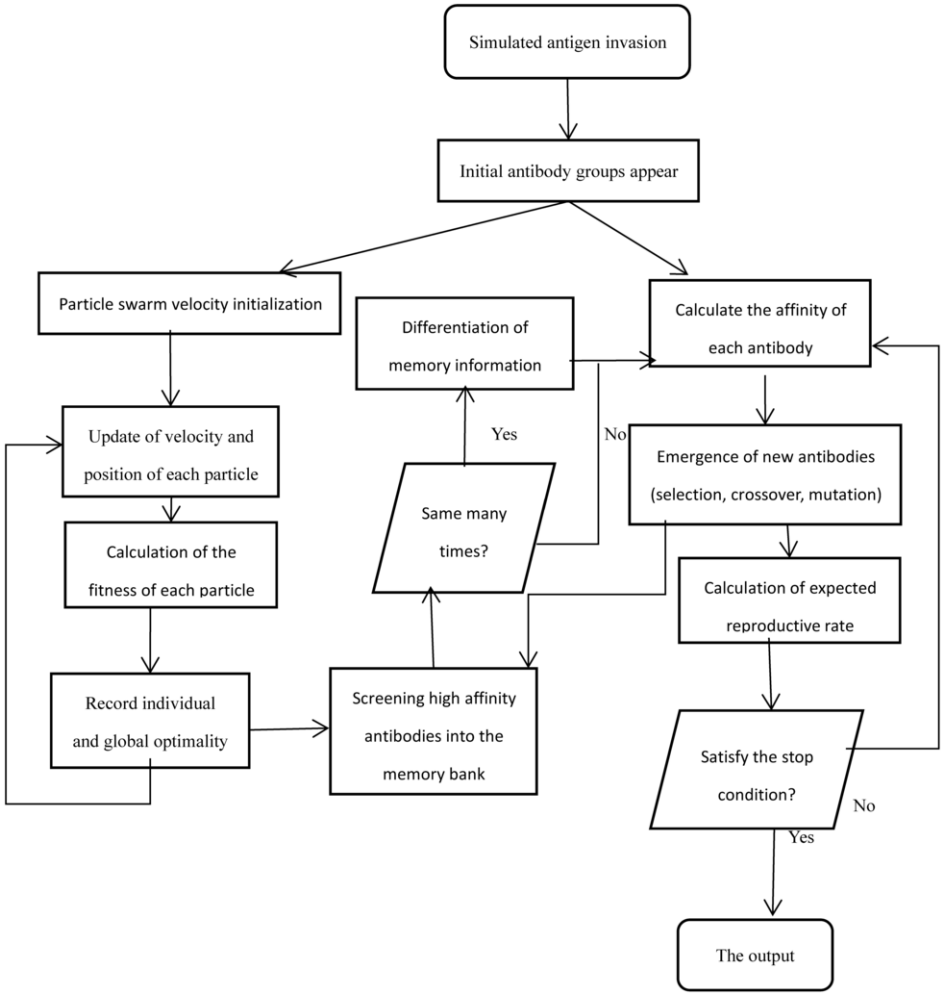

Figure 11. Optimization flowchart

6. Summary and Outlook

This paper mainly studies the problems and applications related to big data security audit based on the computer immune algorithm model, and briefly expounds the relevant concepts of big data security audit, the principle and mechanism of immune algorithm, and the application mechanism and optimization of immune algorithm in big data security audit. The application of immune algorithm model to big data security audit provides a strong technical support for data security protection of large enterprise data system. In view of the optimization of immune algorithm in the application of big data security audit, particle swarm optimization algorithm is adopted for parallel computing, which can improve the speed and accuracy of retrieving abnormal behavior and fault of data system.

Judging from the research results, there are still many problems to be solved. For example, self-adaptive adjustment of parameters can be realized to improve the system self-adaptability; Enrich the collection index, test and verify on more large data systems, and improve the universality and practicability of the application of relevant technologies.

References

[1] Hu Nengpeng, Huang Kunhao, Zheng Lei 2018 Security Audit Based on Big Data [J] *Computer & Telecommunications*, 2018 (10).
[2] Jiangxi2019 Big Data Security Audit Framework and Key Technology Research [J] *Information Security Research* 5(05).
[3] Shi Yong 2021 Exploring the Methods and Contents of Data Security Audit [J] *China Internal Audit*,2021,(02).
[4] Peng Ya 2020 Combined with the data security security audit exploration and practice of "forward" [J] *Digital Communication World*, 2020,(10).
[5] Liu Guocheng 2020 Modeling and Evaluation of Internet Security Audit Process in The Era of Big Data [J]. *Lanzhou Journal*, 2020 (06).
[6] Jiang Yaping, Zhang Ankang, Li Xing 2021 Research Progress and Prospect of Artificial Immune System [J] *Information Security and Communication Security*,2021,(02).
[7] Jiao Jiachen, Bao Nengsheng, Jiang Jiahua 2021 Digital Processing of Ancient Books Based on Artificial Immune Algorithm [J]. *Journal of Shantou University (Natural Science Edition)* 36(01).
[8] Li Ruixue, Ma Liang, Liu Yong 2021 Simulation of Capacity Limited Plant Location Based on Improved Immune Algorithm [J]. *Computer Simulation (Accepted final version)*.
[9] Zhang Yu, Du Mengmeng, Zhang Hongyan, Li Hu 2021 Optimization of Internet of Things System Architecture Based on Immune Evolution [J]. *Application Research of Computers* 38(07).
[10] Sun Fei-yang, Gong Tao 2021 Application of Improved Immune Network and Its Algorithm in Distribution network Fault Location [J] *Modern Computer* 27(23).