



Analysis and Research on Big Data Storage Technology Based on Machine Learning

Xin Li()

Southwest Petroleum University, Chengdu, China
2445498129@qq.com

Abstract. With the advent of the Internet era, the scale and application areas of the Internet are constantly developing, and the Internet has gradually been widely used in people's daily life, economy, military, science and technology, education and other related fields, and its basic and global position and role have become stronger and stronger. Along with the further development of network technology, the issue of network security has become an important factor affecting the development strategy of the country and the economic development of the society. However, in the face of the increasingly complex network structure and large scale, especially the new means of attack using various systems and their security weaknesses, many weaknesses are widely used by intruders, network information systems are facing increasingly serious threats and security risks. Therefore, a variety of techniques based on network security propositions are also changing day by day, and the introduction of advanced concepts such as complex network concepts and genetic algorithms has become a research direction in computer network security evaluation techniques. This paper combines big data-based learning on economic development with a new research and combination of exploration, which is a new research idea and method that lays the foundation for later research.

Keywords: Network Security Assessment · BP Neural Network Algorithm · Genetic Algorithm · Complex Networks

1 Introduction

The advent of the Internet era has allowed the scale of the Internet to expand and its application areas to expand, and the Internet has gradually penetrated into people's daily lives and related fields such as the economy and military and science and technology education, not only allowing its role to gradually increase, but also the foundation and overall picture of its status to be consolidated [1]. Therefore, under the severe network and information process, how to assess the security of the network and use the best solution to reduce or avoid the economic loss caused by the leakage and destruction of information is a strategic issue that needs to be properly addressed [2].

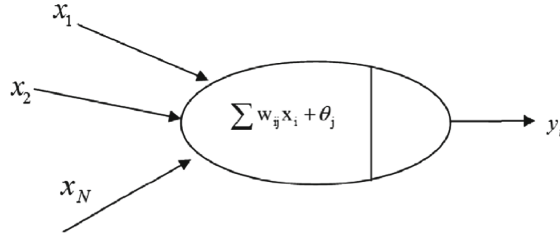


Fig. 1. General description of neurons

2 Neural Network Overview

Neural network (artificial neural network, ANN), is a large scale nonlinear dynamics and parallel distributed information processing system composed of many simple processing units (i.e., neurons or nodes) interconnected by borrowing the characteristics and structure of the human brain [3]. It has a series of characteristics such as huge amount of parallelism and structural variability as well as high non-linearity and self-organization and self-learning. Therefore, it is particularly good at solving problems in consciousness, reasoning, and thinking [4].

3 The Basic Principles and Models of Artificial Neural Networks

3.1 The Basic Principle of Neural Network Composition

(1) Artificial neuron model

A neural network is a parallel distributed system composed of many simple processing units, connected by variable weights [5]. In addition, the basic processing unit of neural network is neuron, and its structure is shown in Fig. 1.

In the figure, x_i is the input signal, w_{ij} denotes the weight of its connection from the i -th neuron to the j -th neuron, and the j -th neuron whose threshold is θ_j . Let s_j be the external input signal and y_j be the output signal, the transformation of the j th neuron in the above model can be described as:

$$y_j = f\left(\sum_i w_{ij}x_i - \theta_j + s_j\right) \quad (1)$$

Here the function $f(x)$ using nonlinearity can be a step function and a segmented function as well as a Sigmoid type function.

3.2 Structure of Neural Networks

In addition to the unit characteristics, the topology of the network is also another important characteristic of the NN. According to the classification of the topology of the network, artificial neural networks can be classified into 3 major types [6].

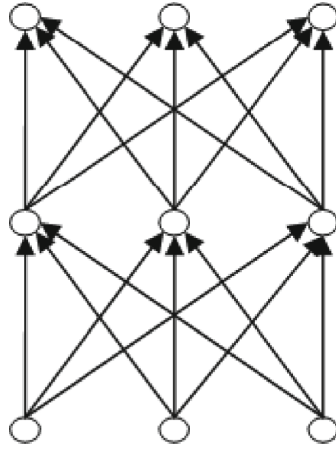


Fig. 2. Feed-forward networks with hidden layers

(1) Interconnected ground network

Any neuron in an interconnected network may be connected to each other, and the information between neurons can be repeatedly transmitted, resulting in a network whose state is constantly changing [7].

(2) Hierarchical feed-forward network

A network in which the neurons are arranged in layers and divided into input and hidden layers and output layers is a hierarchical feedforward network [8]. Each neuronal unit receives input from the upper layer and then outputs it to the lower layer without any feedback (as shown in Fig. 2). Feed-forward networks can be divided into multiple layers, and each layer can only receive input from neurons in the upper layer.

(3) Feedback-based hierarchical network

As shown in Fig. 3, the network is built on the basis of a hierarchical feed-forward network that feeds back the network's output to the network's input, with the feedback being able to feed back both the entire output and a portion of the output [9].

4 BP Algorithm

4.1 Mathematical Description of BP Algorithm

The main idea of the BP algorithm is to divide the learning process into two major stages: first (forward propagation process), the input information is given and the actual output value is processed and calculated for each unit from the input layer to the implicit layer, and second (reverse process), if the desired output value is not obtained at the output layer, the difference between the actual output and the desired output is recursively calculated layer by layer (i.e., error), and the weights are adjusted according to this difference [10].

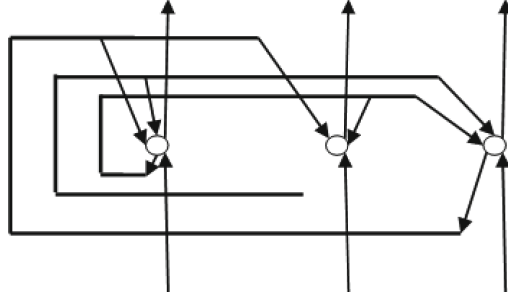


Fig. 3. Single layer fully connected feedback type network

The structure of the multi-layer feedforward network based on the BP algorithm is shown in Fig. 4.

This network has not only nodes in the input layer and nodes in the output layer, but also has one or more layers of implied nodes.

To simplify this process, it is determined that there is only one y output in the network. For N given samples $(x_k, y_k) (k = 1, 2, \dots, N)$, let the output of any node i on the sample be O_i , for a given input as x_k , then the network output is y_k , the output of node i is O_{ik} , The j th cell of the current study layer 1, at the input of the k th sample, the input of node j is:

$$net_{ij}^l = \sum_j w_{ij}^l o_{jk}^{l-1} \quad (2)$$

$$o_{jk}^1 = f(net_{jk}^1) \quad (3)$$

Among them o_{jk}^{l-1} denotes layer 1, and at sample k input, the j th cell whose node output is o_{jk}^{l-1} .

The error function is shown below:

$$E_k = \frac{1}{2} \sum_l (y_{lk} - \bar{y}_{lk})^2 \quad (4)$$

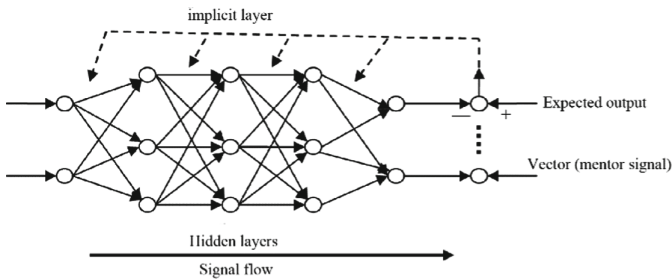


Fig. 4. Multi-layer feed-forward network structure based on BP algorithm

Among them \bar{y}_{lk} is the unit j its actual output. Its total error is shown in Eq. 5.

$$E = \frac{1}{2N} \sum_{k=1}^N E_k \quad (5)$$

Definition:

$$\delta_{jk}^l = \frac{\partial E_k}{\partial net_{jk}^l} \quad (6)$$

Thus:

$$\frac{\partial E_k}{\partial w_{jk}^l} = \frac{\partial E_k}{\partial net_{jk}^l} \frac{\partial net_{jk}^l}{\partial w_{jk}^l} = \frac{\partial E_k}{\partial net_{jk}^l} o_{jk}^{l-1} = \delta_{jk}^l o_{jk}^{l-1} \quad (7)$$

In this training sample its presentation order should be randomly generated from one round to another. The parameters of momentum and learning rate should be adjusted as the number of training iterations increases.

4.2 Shortcomings of BP Algorithm

The neural network in BP algorithm is a feed-forward network in terms of information flow during operation. This network not only provides many neurons with simple processing power, but also has a complex and nonlinear mapping capability without feedback, so it is not a nonlinear dynamical system. It can only be considered as a nonlinear mapping. However, it is still very important because of its wide applicability and completeness in theory, but it also has number of problems.

- (1) The BP algorithm converges according to the direction of the mean square error of its gradient descent, however, the mean square error of its gradient curve still has a number of global and local minima, which causes the neural network to fall into the local minima very easily;
- (2) BP learning algorithms whose convergence speed is very slow, which may lead to a lot of wasted time;
- (3) The number of nodes implied by the network and its selection currently lacks a complete and unified theory for guidance.
- (4) The learned network has a poor generalization ability. In view of the above problems, the basic BP algorithm needs to be improved as necessary to speed up the convergence speed and achieve optimization.

The flowchart of the genetic algorithm is shown in Fig. 5, and it is oriented to all search problems as a general algorithm for solving search problems.

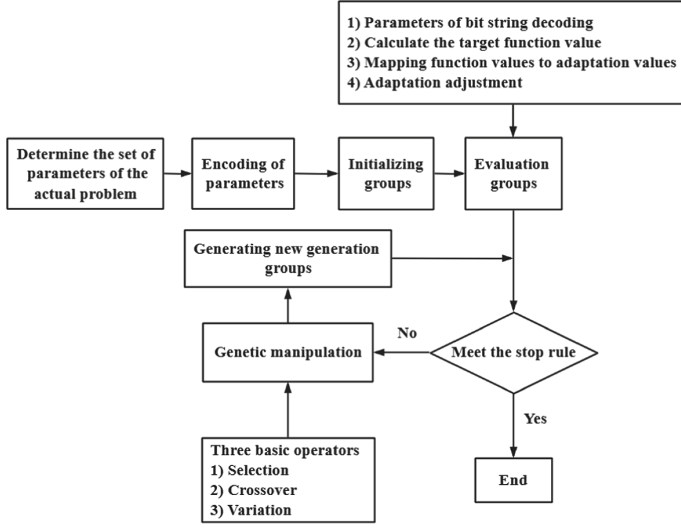


Fig. 5. Genetic algorithm flow chart

5 Improved Modeling of Neural Network BP Algorithm Using Genetic Algorithm (GA)

The mathematical description of the problem of optimizing genetic-neural networks is mainly shown as follows

$$\begin{cases} \min E(w, v, \theta, r) = \frac{1}{2} \sum_{k=1}^{N_1} \sum_{t=1}^n [y_k(t) - \hat{y}_k(t)]^2 \\ s.t. w \in R^{m \times p}, v \in R^{p \times n}, \theta \in R^p, r \in R^n \end{cases} \quad (8)$$

In the above equation $y_k(t)$, $\hat{y}_k(t)$ the above quadratic nonlinear optimization problem is solved by genetic algorithm, and the obtained network its connection rights and structure, calculate E_2 If E_2 less than the set error ε_2 then it is possible to apply the model to the actual forecast. Therefore, this paper uses two different methods to optimize the three-layer neural network of BP, and the main steps of method 1 should be as follows:

Since the genetic algorithm uses the maximum value of the objective function as a function of its fitness in the process of optimization, the fitness function is defined as:

$$F(w, v, \theta, r) = \frac{1}{\sqrt{\sum_{k=1}^{N_1} \sum_{t=1}^n [y_k(t) - \hat{y}_k(t)]^2}} \quad (9)$$

Then Eq. (8) will become as follows:

$$\begin{cases} \max F(w, v, \theta, r) \\ s.t. w \in R^{m \times p}, v \in R^{p \times n}, \theta \in R^p, r \in R^n \end{cases} \quad (10)$$

First, the basic solution space is encoded, and the code string generated by the encoding is composed of two parts: the control code and the weighted coefficients. The

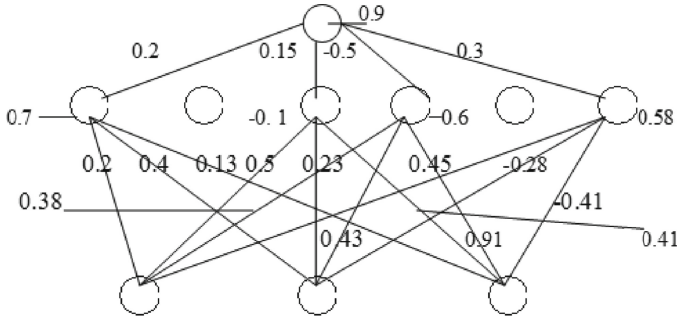


Fig. 6. The structure of the network and the coefficient of the network and its connection rights

control code is used to control the number of hidden nodes by a string of 0–1, where 0 means no connection and 1 means connected, and the string length l_1 can be determined by multiplying the number of input nodes by 0.5 to 1.5. The coefficients of the weights are mainly used to control the connection rights of the network, and the encoding of the floating point number is used, and the string length $l_2 = m \times l_1 + l_1 + l_1 \times n + n$ (here m is the number of input nodes and n is the number of output nodes). The codes are linked into long strings according to some sequential level, each of which corresponds to a set of network connection rights and network structure. We take the node with 3 inputs as an example word, then the hidden layer node will have at most 6 points. A graphical representation of this network structure is shown in Fig. 6.

In accordance with the threshold values and weights marked in Fig. 6 also the network whose connection is able to give the corresponding weight coefficients of the coding string with the coding string of the control.

The control code string is: 1 0 1 1 0 1.

The input layer of the neuron to the hidden layer whose weight matrix is:

$$\begin{bmatrix} 0.20 & 0.50 & 0.38 & 0.45 \\ 0.40 & 0.23 & 0.43 & 0.28 \\ 0.13 & 0.41 & 0.91 & -0.41 \end{bmatrix}$$
 expand this according to some order to become: 0.20 0.50 0.38 0.45 0.40 0.23 0.43 0.28 0.13 0.41 0.91 -0.41 gene strings.

Neurons in the hidden layer have a threshold value of 0.58 -0.60 -0.10 0.70.

Hidden layer to the output layer whose weights are: 0.30 -0.50 0.15 0.20.

The output neuron has a threshold value of: 0.9.

Then the cascade of weight codes and control codes results in a code string of:

1 0 1 1 1 0 1 0.45 0.38 0.50 0.20 0.28 0.43 0.23 0.40 0.41 0.13 0.41- 0.91 0.58 0.60- 0.10- 0.70 0.30 0.50- 0.15 0.20 0.9

The probability of P_c is used to crossover the selected individuals with each other. Let a mutual crossover be performed between the i -th individual and the $i + 1$ -th individual, and the crossover operator will be as follows.

$$\begin{cases} X_i^{t+1} = c_i \cdot X_i^t + (1 - c_i) \cdot X_{i+1}^t \\ X_{i+1}^{t+1} = (1 - c_i) \cdot X_i^t + c_i \cdot X_{i+1}^t \end{cases} \quad (11)$$

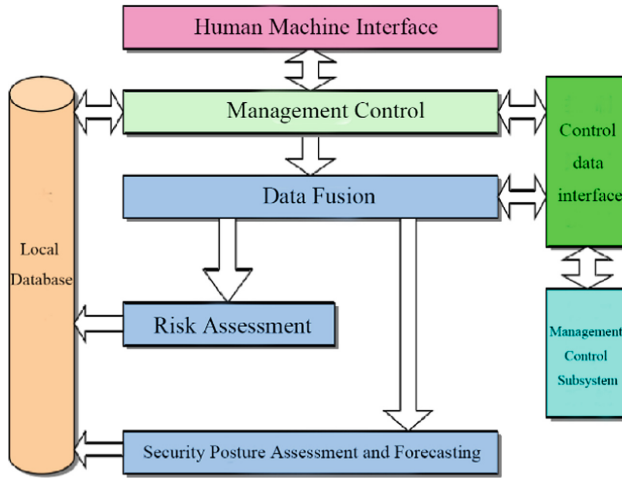


Fig. 7. Assessment of security posture and prediction of subsystems and their composition

In the equation X_i^t, X_{i+1}^t is a pair of individuals prior to crossover; X_i^{t+1}, X_{i+1}^{t+1} then the individuals after the crossover; c_i is a uniformly distributed random number located in the interval $[0,1]$.

6 Design of the System

6.1 System Composition

The human-machine interface, management control, data fusion, risk assessment, security posture assessment and prediction, and local database modules in this subsystem make up the system, as shown in Fig. 7.

7 Conclusions

This paper firstly introduces the main contents of network security assessment design from a general perspective, mainly from three aspects, including the design of support platform architecture, the design of network security assessment module composition and workflow, and the design of system functional modules. This paper mainly focuses on the assessment of network risks and elaborates on the contents involved in the assessment process of network security, and then uses it as a basis to study the situational awareness of network security. We also study the support platform for security testing and evaluation in this system based on the above research, and carry out detailed design and implementation of the subsystem for evaluation and prediction of security posture. A subsystem for security posture assessment and prediction is implemented, and experiments are conducted with this subsystem in conjunction with other subsystems in the support platform, and then a small network is comprehensively tested and evaluated, and eventually a variety of manifestations are adopted, and a trend graph of the system's

security status is visually displayed to the user. In the network posture prediction, three prediction models are used in this system, and the prediction results are not satisfactory when the number of evaluations is small, so further research on the prediction methods of these models with less historical data is needed to make the prediction results reflect the approximate posture direction of the network.

References

1. Ma Xiaofeng. Exploring the standardized use of big data technology in network security analysis[J]. China Standardization, 2022(08):11–13.
2. Song YJ. Discussion of computer information processing technology in the context of big data[J]. Information Record Materials, 2021, 22(06):123–125. <https://doi.org/10.16009/j.cnki.cn13-1295/tq.2021.06.068>.
3. Luo Shanshan, Yang Fang, Ren Wei. Exploration on the development of management accounting in the context of big data[J]. Business Accounting, 2021(08):102–105.
4. Jia Yan. Systematic study of computer application technology in the context of “Internet+”[J]. Digital Technology and Applications, 2020, 38(08):163–164. <https://doi.org/10.19695/j.cnki.cn12-1369.2020.08.61>.
5. Yu E. Cheng. The application of computer software technology in the era of big data[J]. Electronic Testing, 2020(15): 139–140. <https://doi.org/10.16520/j.cnki.1000-8519.2020.15.057>.
6. Peng Junzhen. Analysis of energy-efficient storage strategy for remote sensing images based on Hadoop[J]. Electronic Technology and Software Engineering, 2020(13):182–183.
7. Yang Ling. Research on innovation of local government governance under the perspective of big data[D]. Southwest University, 2020. <https://doi.org/10.27684/d.cnki.gxndx.2020.002501>.
8. Wei, Jenna. Research on computer information processing technology in the era of big data[J]. Hubei Agricultural Mechanization, 2020(02):158.
9. Cao J. Research on big data storage security technology in the context of cloud computing[J]. Information Systems Engineering, 2020(01):51–52.
10. Tang Mingshuang. The use of computer network security technology in the era of big data [J]. Heilongjiang Science, 2019, 10(24):108–109.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

