Original Article

Identification of Defect Location in Computer Networks Using Genetic Algorithm and Artificial Neural Networks

Mahyar Najizad

Software Engineer

Abstract

Localization is the basic and fundamental principle of managing computer networks. The first step in each network is to detect intrusions in computer systems and then identify and determine the defective and corrupted components. In this paper, Growing based on the identification of location and the exact location of defects in computer systems by the method of genetic algorithm is presented. In the first step, the method of the genetic algorithm is described and then, by forming the initial population, we investigate the location of the defects. Then, the artificial neural network method determines the error and failure rate at the designated location. The results of the study show that the proposed method based on genetic algorithm and artificial neural network minimizes the cost of testing and locating defective components in relation to existing ones and has a better performance than similar methods.

Key words: Troubleshooting, Genetic algorithm, Artificial neural networks, Computer networks

INTRODUCTION

With increased speed, efficiency, number and relevance of computers in the 1970s, the need for security systems has grown a lot. In 1977 and 1978, the International Standards Organization convened a meeting between governments and EDP inspection bodies, resulting in a report on security, inspection and control systems at that time. At the same time, the US Department of Energy began a very rigorous investigation into the safety and security of computer systems due to concerns about the security of its systems [1]. Between 1984 and 1986, they conducted research on the security of computer systems, which resulted in the production of a real-time intrusion detection system that operated on the basis of expert systems. This system was named IDES. In this project, a combination of anomaly detection and abuse detection was studied. The idea presented in this project was used as the basis for many of the intrusion detection systems that were developed since then. On the other hand, after intrusion detection, corrupted components in computer networks need to be detected, located and repaired to ensure network health. There are several reasons why components of a computer network are corrupted; for example, interrupting links or failing nodes are one of the reasons that lead to network failure. Failure in the network has a negative effect on network performance, resulting in loss of packets and loss of communication paths. Defective localization is considered to be the main aspect of defect management in computer networks [2]. Given the presence of corrupted components in computer networks is inevitable, it is imperative that the defect management department immediately locates defective components. The most important point in troubleshooting is the cost of testing the components of the computer network. Given that the cost of testing different components of the network to find defective components is expensive, then components should be chosen for testing that are more likely to be defective to other components.

Now, how to determine these probabilities and which component to test, so that the cost of finding the defect

Corresponding Author: Mahyar Najizad, Software Engineer. Phone:+989301649865. E-mail: Najizad1373@gmail.com

is minimized, is NP-Complete issue. Most intrusion detection systems functioned as host-based, which means that they gathered information from the operating system or application and analyzed them [13]. Therefore, providing an algorithm that can accurately identify the defect location with the least possible complexity is of great importance [2-3]. Therefore, providing an algorithm that can accurately identify the defect location with the least possible complexity is very important. Various methods are available for diagnosis and troubleshooting in computer networks. since creating computer systems (hardware and software) without weaknesses and security failures is impossible technically, the intrusion detection in computer system research is of particular importance [13-14]. In the research, an algorithm was developed based on an active strategy, which employs a number of packets for defect detection on a part of the network at any moment that reduces the additional traffic overhead at the diagnostic stage [1]. Heuristic algorithm has been presented based on passive strategy. This method selects the components for testing that have the most common use in different paths, in addition to the fact that additional traffic is not sent to the network, and the test cost is also minimized [2]. Further, researchers have developed a strategy based on an active strategy that divides the defect diagnosis procedure into multiple stages, at each stage only performs diagnostic operations in a small part of the network using a few extra packets. This method for defect diagnosis tries to have less effect on the network traffic [2]. On the other hand, they have introduced a heuristic algorithm for selecting candidate components to test these components that minimize the cost of testing and localization [3].

Intrusion

Intrusion is called to the collection of illegal actions that compromise the integrity or security of a source. Intrusions can be divided into internal and external categories. External Intrusions are those Intrusions that are committed by authorized or unauthorized persons from the outside of the network to the internal network, and internal Intrusions are carried out by authorized persons on the system and the internal network from within the network itself.

Intrusion Methods
• Software defects
• Breaking passwords
• Eavesdropping on network traffic
• Design weaknesses in the network
• Services
• Network computers

Intrusion Detection Methods
In order to deal with the attackers of computer systems and networks, several methods have been developed called "Intrusion detection methods" that monitor the events occurring in a system or computer network. Given that corrupted components are inevitable in computer networks, it is imperative that the defect management department immediately localizes defective components [4]. The most important point in troubleshooting is the cost of testing the components of the computer network. Intrusion detection methods are divided into two categories:
• Abnormal behavior diagnostic method
• Abuse detection method or signature-based diagnosis

In an abnormal behavior diagnosis method, one facet of normal behavior is created. An abnormality may indicate an Intrusion to identify abnormal behavior, and should identify normal behaviors, and find patterns and rules for them. Behaviors that follow these patterns are normal, and events with a more than standard deviation of these patterns are identified as abnormal behavior. Abnormal Intrusions are very difficult to diagnose, since there is no steady pattern for monitoring. Usually an event that occurs more or less than two standard deviations from normal statistics is assumed to be abnormal. In the method of abuse detection or signature-based detection, this technique emulates attacks, and compares that event with existing events in database when suspicious events are replicated and declares intrusion If matched. Host-based intrusion detection systems are usually software that is installed on the computer. These intrusion detection systems use different methods to detect an attack. One of the easiest things to do is to ensure that system files are accessible and safe. These systems work well in the files section. Detecting whether the sensitive files of the system have been manipulated or not and/or tracking unauthorized access by a user outside of the work area and the specified access are the functionality of the system.

Defect detection strategies in computer networks are categorized as active or passive. In the active methods, a series of additional information and packets are sent to the network to detect the corrupted network components. Given that these packets and information are sent only to detect the corrupted network components, so network traffic increases. Inactive strategies use end-to-end data on the network to localize defective components [5]. The advantage of a passive strategy to an active strategy is that it does not send additional traffic to the network, and thus the network traffic does not increase, but in the active strategy, measurements for defect detection are often more accurate than the passive strategy. When the network is small in scale and has a simple topology, the active strategy works very efficiently. But when the scale of the network gets bigger and its topology becomes complex, the active strategy may not be feasible and will consume a lot of bandwidth.

Host-based Intrusion Detection Systems

Host-based intrusion detection systems are usually software that is installed on the computer. These intrusion detection systems use different methods to detect an attack.

One of the easiest things to do is to ensure that system files are accessible and safe. These systems work well in the files section. Detecting whether the sensitive files of the system have been manipulated or not and/or tracking unauthorized access by a user outside of the work area and the specified access are the functionality of the system.

Genetic Algorithm

In terms of this technique, the intrusion detection process involves definition of a vector for information of events, that is, the corresponding event indicates that the event is an intrusion or delay. The mechanisms of this operation are such that at the start of the process a hypothetical vector is considered and its validity is considered, then another assumption is made, which is based on the results of the test of the previous assumption. This is done by the genetic algorithm. The production of the new assumption is repeated repeatedly so that the solution can be determined.

Genetic Algorithm Data

Intersecting operators

This operator is for generating new chromosomes in the genetic algorithm. An intersection operator similar to its counterpart in nature produces new people whose genes are generated from their parents. In this paper, we have chosen the operator in terms of which operator provides a more appropriate response.

Mutation operator

The mutation is a random process in which the contents of a gene with another gene are replaced to produce a new genetic structure. The Gaussian mutation operator is the operator, which is chosen by changing its desired ratios [6]. The variance of this explanation is regulated by scale and Shrink parameters, which is set in this research with successive changes of these variables.

Chromosome

The basis of each function of the genetic algorithm is to convert each set of responses to a coding. In fact, the encrypted form is the probable answer to the problem. In this case, each chromosome is a solution to a problem that may be justified or unjustified.

Fitness Function

The value of the problem variable is placed in this function. Accordingly, it can specify the utility of each function, and thus the desirability of each response is determined. In optimization problems, the objective function is used as a fitness function. The suitability of response or not is determined by the amount obtained from the fitness function. There are several ways to apply the model's limitations. Accordingly, the strategy used is based on Equation (1).

$$\overline{x} \notin M$$

$$\tilde{n}(\overline{x},t) = \left\{ \max\{f(\overline{x})\} - \min\{f(\overline{x})\}, \sum_{j=1}^{P} f_j(\overline{x}) \right\}$$

(1)

Population

One of the features of the genetic algorithm, instead of focusing on a point or a chromosome, is to set a set of responses or chromosomes. The model in this study will be solved based on the initial population of 1000 chromosomes.

Roulette Selection Method

Accordingly, the selection function for the new generation uses all babies and parents, or each baby immediately replaces their parents. In this rotation, the rotational level is divided into parts that are equal to the number of members of the current community. The level of each section is proportional to the value of the desired fitness solution. Accordingly, the probability of chromosome selection is obtained by equation (2).

$$p_k = \frac{f_k}{\sum_{i=1}^{N} f_i}$$

(2)

In this equation, f_x is the fitness rate of chromosome k. Accordingly, we use a simulation to implement the roulette wheel model as follows. First, the whole population is converted into a chromosome.

Then, the cumulative probability value for each answer (q_k) for each answer is calculated from the following equation.

$$q_k = \sum_{i=1}^{k} p_i$$

(3)

By creating a random number H at a distance, the index k is determined by the equation (4).

$$q_k = \min\{j : q_j \geq H\}$$

(4)

Accordingly, the smallest index that has a higher cumulative probability of H is selected.

Finally, the weight of the matrix and the parameters specified in Table 1 are expressed. These weights are

Table 1: Specifications of artificial neural network used in simulation

Learning rate	Learning layers	Neuron No	Hidden layers	Input method
0.001	440	20	10	Two-layer *feed-forward*

obtained by taking into account the first stage, without the effect of weights relative to each other, and taking into account the second stage, taking into account weights relative to each other.

Multivariate Linear Regression Model

The matrix regression model can be represented as Equation (5):

$$Y = x\beta + \varepsilon \tag{5}$$

β is Regression coefficient matrix, e is the matrix of fit error and Y is the response matrix. By solving Eq. (1) we have the following equation (6) in terms of β:

$$\beta = (X^{-1} X)^{-1} (X' Y) \tag{6}$$

In this context, is the transpose of matrix X. To calculate the inverse $(X^{-1} X)$ of independent variables, there is not much correlation, as this increases the error due to the rounding of data and computations. To resolve this problem, the correlation between independent variables must be eliminated before the regression model is made. One of the basic solutions for solving this problem is the implementation of the analysis of the main components of the main input variables based on inflation variance factor. In this paper, after solving the correlation problem in independent variables, a suitable model using multiple linear regression Variable is presented [7]. Accordingly, the input of variables into regression varies from one variable to another, from the most important to the least important one.

Artificial Neural Networks

The creation of an artificial intelligence network, which has the learning, creativity and resilience ability, is the main focus of research on artificial intelligence. The network studied in this study is a multi-layer perceptron network, which is a network of feed-forward networks. In this type of network, information goes from one side to the next (through the input layer to the hidden layer and reaches the output layer through the hidden layer).The learning algorithm of this network is an error propagation type, and the type of stimulation function is also Sigmoidal Tangent [8]. Neural networks consist of a series of layers consisting of simple components called neurons that act
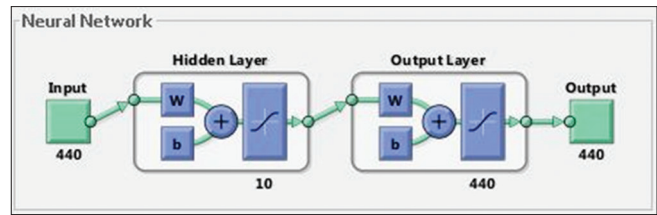
Figure 1: Two-layer feed-forward neuron network model
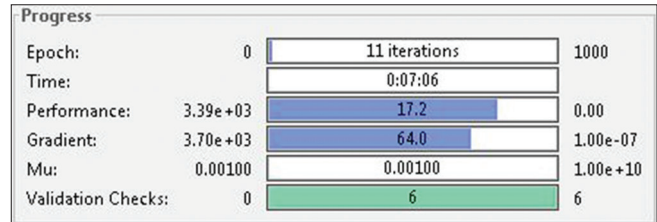
Figure 2: Two-layer feed-forward Neuron network analysis

in parallel. In Figure 2, a simple neuron with input R is shown in which each input vector is weighed by the proper choice of weight W and the sum of bias inputs constitute the input of the moving function.

The most important point in the neural networks is to normalize the input parameters obtained through equation(6). If we consider the input parameters in this system as a signal, then we can claim that the behavior of this signal at time t can depend on the behavior of the signal at time t-1. The input parameters to the network are entered into multi-layer perceptron neural network based on all commonly used methods and 70% of the data for the training phase, 75% for the test phase and 10% for the validation phase have been selected. For input data, 440 items are considered [9].

$$x' = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{7}$$

In this equation, is the input parameter, is the minimum value of parameter x, is the maximum value of the parameter x, is the normalized value of parameter .

The error detection and predictive indicators used in this study are as follows.

The squared mean squared error value

A criterion for obtaining the best estimate

RMSE value is determined based on *mean squared error (MSE)*.[12].

$$RMSE = \frac{\sum_{i=1}^{n} (y_{est} - y_{act})^2}{n} \tag{8}$$

When the network parameters are obtained after a complete period of patterns presentation, in the phrase epoxy or a cycle is called to this type of repetition; the number of network replicators is equal to the number of learning data.

Problem Statement

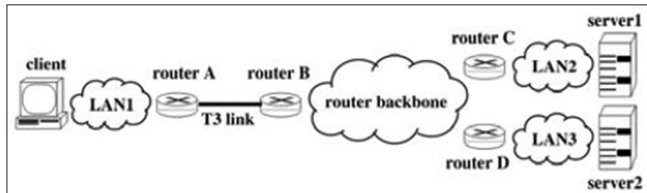Figure (3) shows the various components of a computer network with different protocol layers. The failure of each

Figure 3: An example of a network that contains different components in different protocol layers
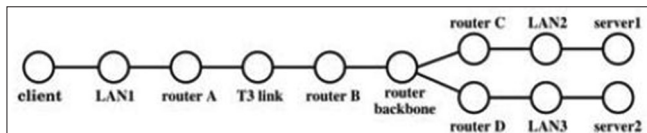

Figure. 4: Logical topology related to Figure 3

of these components makes the connection between the client and server difficult. Localization in such a network seems difficult. Therefore, in the field of computer networks, we need a way to identify and resolve defective components at the network level. Accordingly, the physical components of the network and the physical topology of the network should be converted into a logical topology. Thus, the form represented in Figure. 3 can be represented as in the protocol of Figure. 4.

For simpler management of network and study and identification of defect location in different layers, we use end-to-end data in the network. With regard to end-to-end data, it can detect the behavior of paths in the network. If any path shows an abnormal behavior in the delivery of data, it indicates that there are one or more defective components in this path. Using the information available at the endpoint (server), you can detect broken paths.

According to the explanations provided, the corrupted path can be obtained, but the type of node, the cause of the failure and localization of the node cannot be gained. To do this, you need to test a sequence of nodes in the
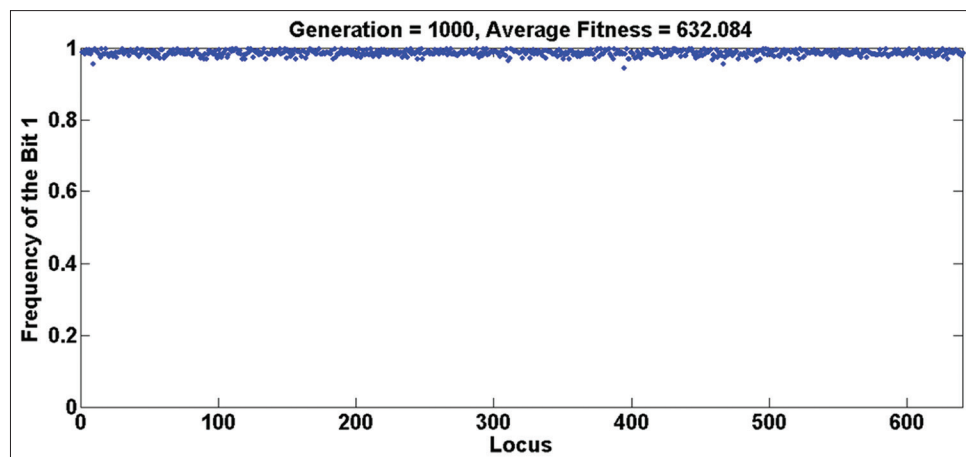

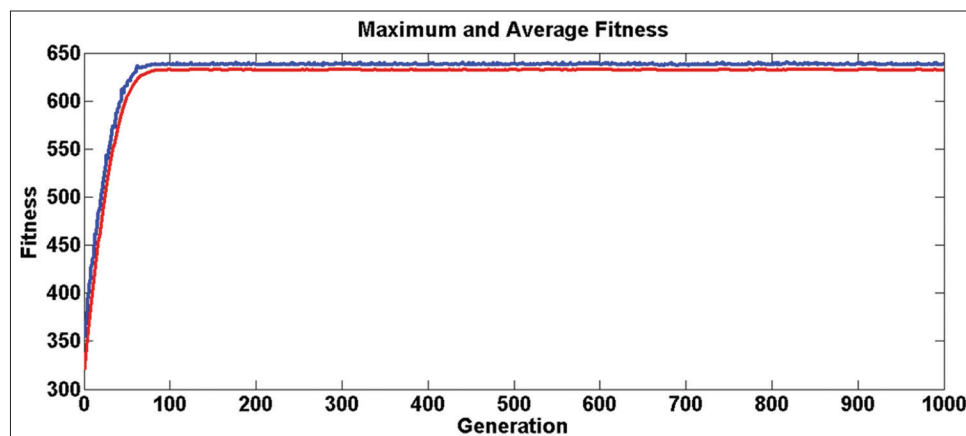**Figure 5: Analysis of the presented features in terms of the number of chromosomes in 1000 and the specified weights**

Figure. 6 : Maximum value and average value of Fitness function with respect to the weights specified in the genetic algorithm.

path to be repaired in case of failure. Finding and locating defective nodes also have separate costs that this cost can be reduced to some extent by the methods outlined in this article. Accordingly, we have potential information from the network paths (that is, we are informed of the nodes that each client uses to transfer data to the server, and we do not know exactly which paths each client uses to transfer data to the server. So we call them a client-server pair, which we use word 'pair' for a simple. So, we can know receive rate between client and server through the end-to-end data. If n packets of client are sent to a server in a network, m packets should reach to destination successfully under healthy network. In this case, receive ratio of m/n is obtained.

Evaluation of Network Performance
One of the basic criteria for the validity of the results obtained from the linear regression model and artificial neural network

is the fitness model of the correlation coefficient (R), whose value is between (-1) and (1), and the absolute value close to (1) displays better matching of estimated data (the number of healthier nodes). In this research, for the analysis of the results, the amount of layers in this method is determined according to Table 1 and the values of the error indices for the three stages of training, testing and validation are established according to Table 2 [12].

Analysis Method
The analytical charter is based on the proposed method, as follows:

At first, a threshold is considered in a computer network.

If the paired receive ratio is less than this threshold, there are at least one defective node in this pair and that pair is

Table 2: The amount of error indicators for the three stages of training, testing and validation in the simulation

RMSE-Train	RMSE-Test	RMSE- Validation	NDEI-Train	NDEI-Test	NDEI-Validation
0.08963	0.0023	0.0658	0.5489	0.02456	0.61426
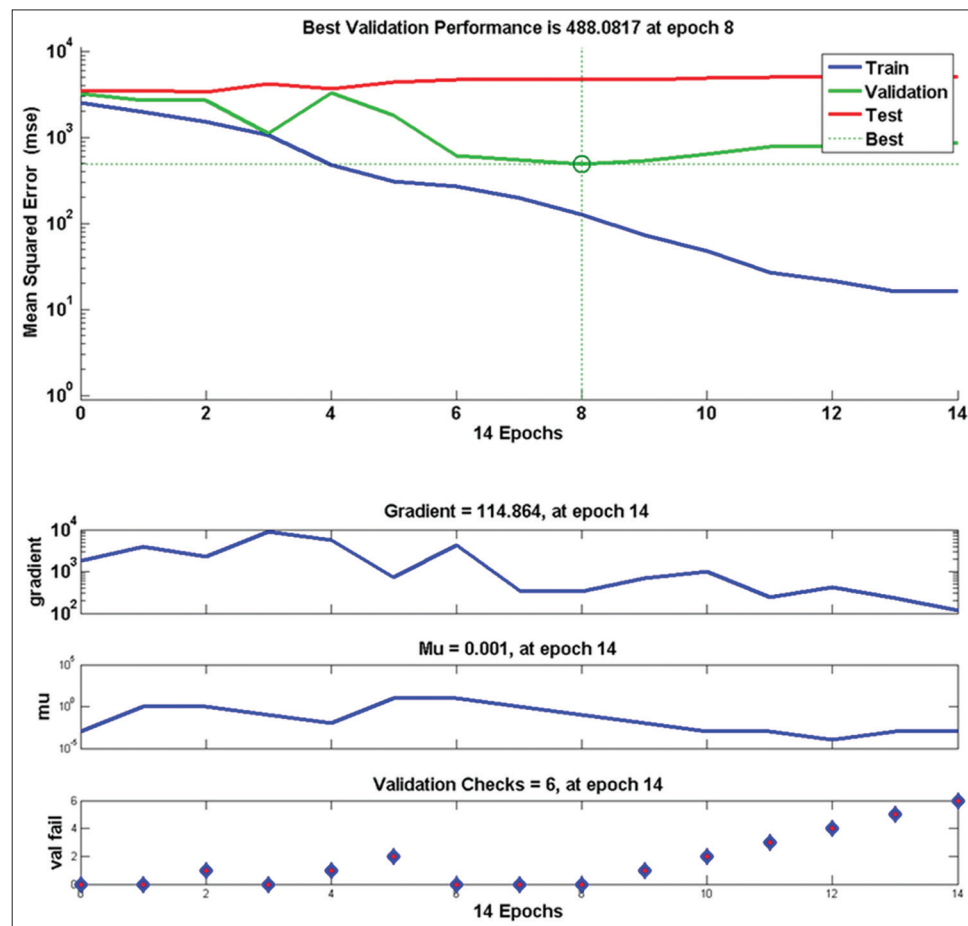
Figure 7: The accuracy of the finding the defective chromosome, the variation in the cost of the test and the number of healthy chromosomes relative to the defective chromosomes in the first scenario

known as a corrupted pair, otherwise the pair is known as a healthy pair. In the event that at least one healthy node cannot be detected by end-to-end data, a node is randomly selected, If this node is healthy, The algorithm enters the next step, otherwise a defective node will be selected and repaired and the algorithm enters the next phase.

In the next step, the difference between the healthy and corrupted nodes is determined. To determine the difference between healthy and corrupt nodes, the following is used:

$$D_i = bad\ pair_i - \left(Safe\ Nodes \cap bad\ pair_i \right) \tag{9}$$

In this equation, is the difference between the corrupted and the healthy nodes; is meant ith corrupted pair. is the set of healthy nodes identified is HAND. This phase of the genetic algorithm method is used to form nodes. After applying genetic operators, the whole created new population has been replaced with the whole previous population. With the creation of new chromosomes, two conditions may occur. $|D_i|$, In which case the difference between the healthy and defective nodes reaches a node. So in this case if the selected node is defective, the node is added to the list of repair nodes and the algorithm is repeated from the first stage. If the node is not defective, the node will be added to
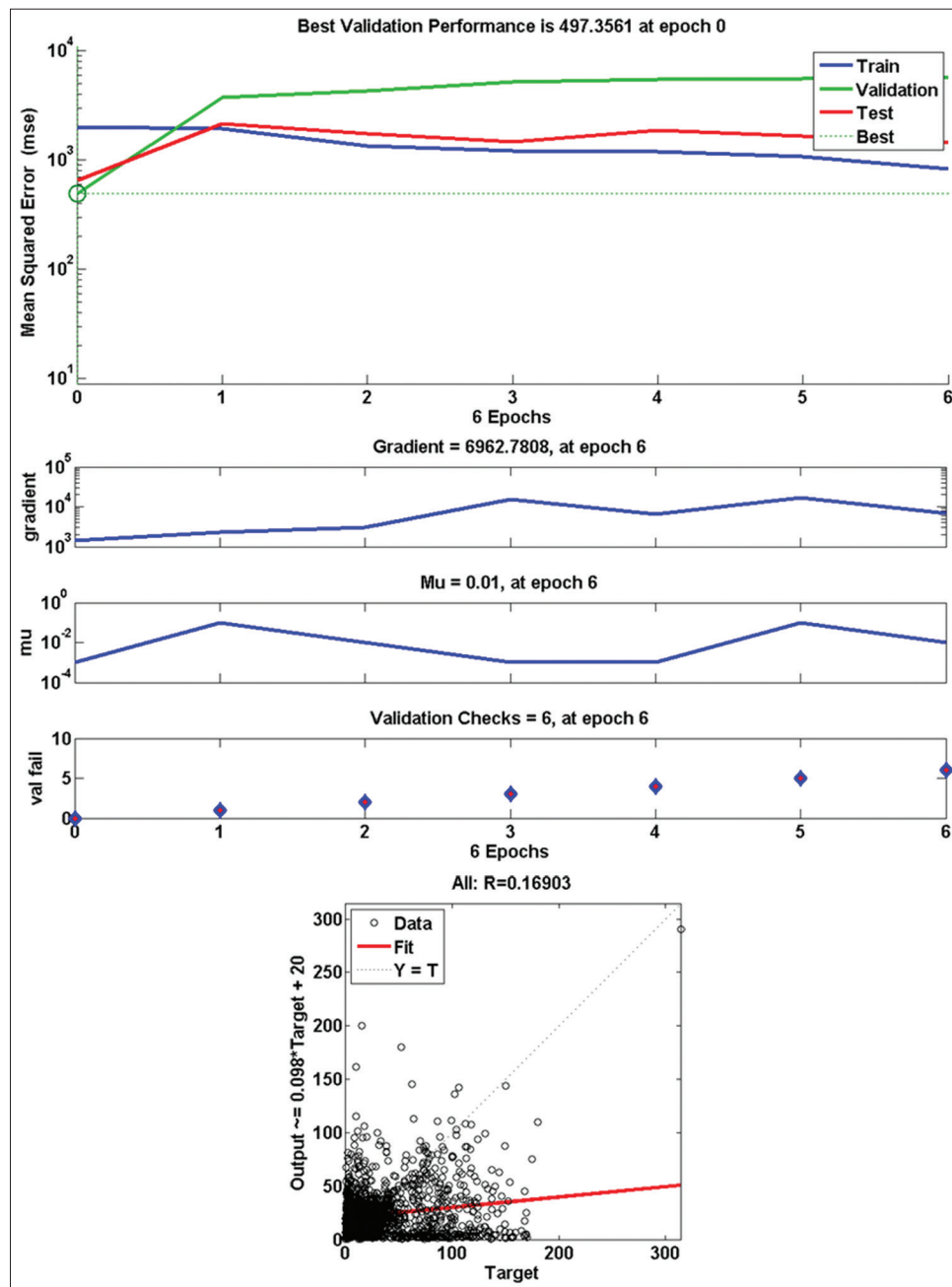

Figure 8: The accuracy of the finding the defective chromosome, the variation in the cost of the test and the number of healthy chromosomes in comparison to the defective chromosomes in the second scenario

the list of healthy nodes and the next chromosome will be formed. If $|D_i|>1$, in this situation the next node (the next chromosome) is checked. If the difference in chromosomes is greater than the (healthy and the next chromosome) than one, then the genetic algorithm is repeated. Otherwise, the algorithm will be repeated until all defective chromosomes are detected. In this simulation, it is supposed that in certain periods of time, information of pairs is sent to the server via a reporter service. It is also assumed that the cost of testing the network components is homogeneous, i.e. the cost of testing each component in the network is equal. Therefore, in

simulations, the cost of testing each component is considered as one unit. In addition, we do not know the probability of defective network nodes, so we consider the probability of all nodes equal. Table 3 illustrates the scenarios used in this simulation. The number of chromosomes processed for analysis in the genetic algorithm is 1,000. The analysis of the presented features is shown according to the weights indicated in Figure. 1.

Figuers 7, Figure 8, and Figure 9 show the accuracy of finding the defective chromosome, the cost changes related
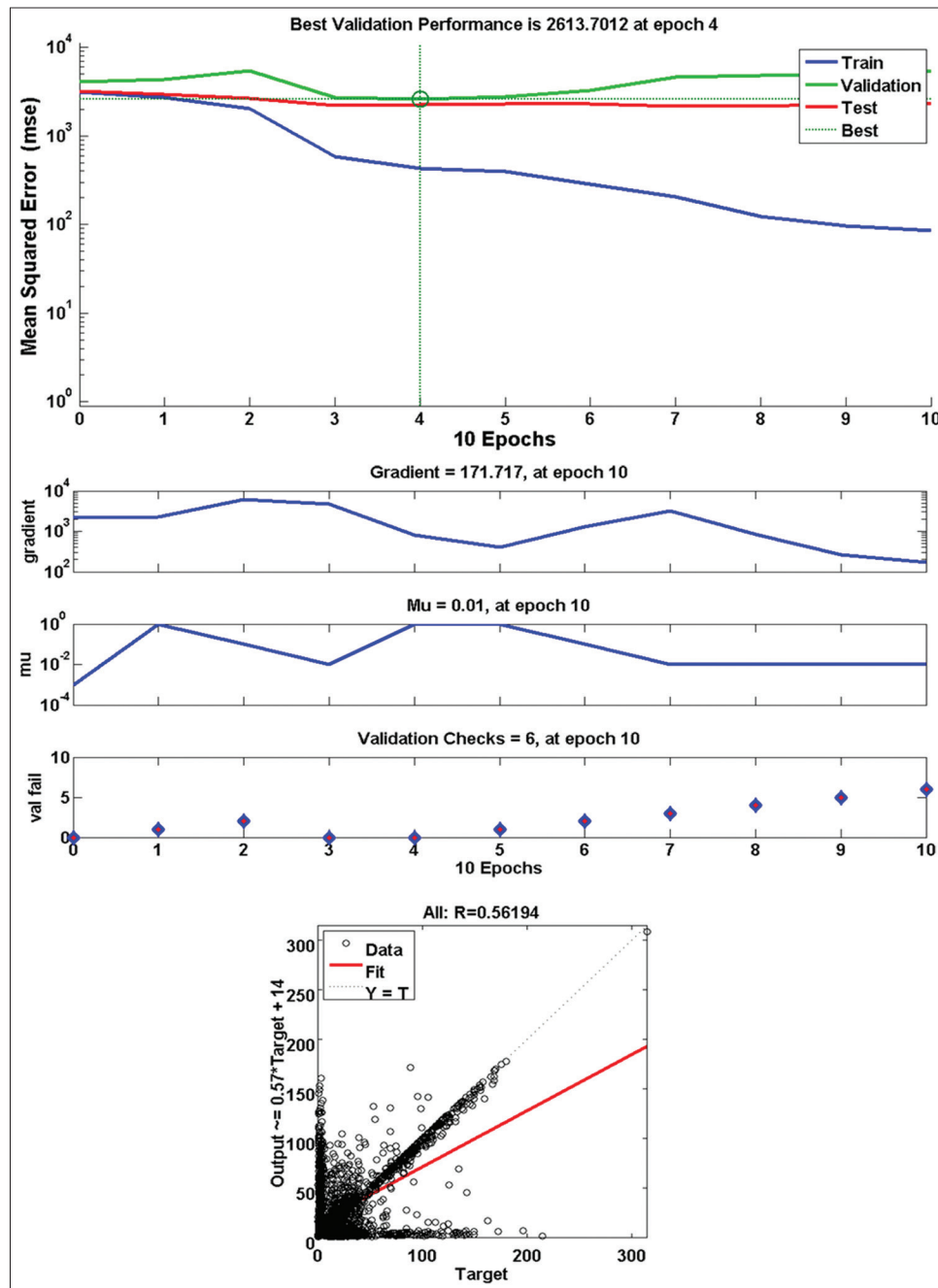

Figure 9: The accuracy of the finding the defective chromosome, the cost variation of the test and the number of healthy chromosomes in comparison to the defective chromosomes in the third scenario

Table 3: Parameters for the considered scenarios

scenarios	F	C	N
Scenario 1	0.1	1-100	80
Scenario 2	0.1	100-250	160
Scenario 3	0.1	250-1000	250

to the test, and the number of healthy chromosomes relative to the defective chromosomes, respectively. Further, based on these results, the larger the number of clients, the total cost of testing also decreases. So the ratio of sum of testing cost to the number of clients is the inverse ratio.

The inverse ratio is due to the fact that as the number of clients increases, there will be a large number of pairs in the network, so more healthy nodes can be detected in the network, so the more informed of status of healthy nodes, it can guess the location of defective nodes and test them Better and more accurately. Therefore, the total test cost decreases with increasing number of clients.

CONCLUSION

In this paper, a sequential testing problem to find defect in computer networks was defined. Then, an algorithm was proposed for finding errors in computer networks that detects errors in the network through end-to-end data and then locates the defective components using the genetic algorithm. The results of simulations in different scenarios show that the proposed method finds error with less cost than existing methods. On average, the proposed method at 0.96% of cases will select the right component for testing. Therefore, based on the obtained results, the proposed algorithm is a suitable method for finding defects in computer networks.

REFERENCES

[1] Mohamed and O. Basir, "A New Probing Scheme for Fault Detection and Identification", IEEE International Conference on Electro/Information Technology, 2009.

[2] B. Wang, W. Wei, H. Dinh and W. Zeng, "Fault Localization Using Passive End-to-End Measurements and Sequential Testing for Wireless Sensor Networks", IEEE Transactions on MobileComputing, 2012.

[3] Y. Yu, Q. Liu and L. Tan, "A graph-based proactive fault identification approach in computer networks", Elsevier, Computer Communications, 2005.

[4] L. Lu,Z. Xu, W. Wang and Y. Sun, "A new fault detection method for computer networks", Elsevier, Reliability Engineering and System Safety, 2013.

[5] M. Steinder and A. Sethi, "A survey of fault localization techniques in computer networks", Elsevier, Science of Computer Programming, 2004.

[6] Khorasani, S. T., & Almasifard, M. (2017). Evolution of Management Theory within 20 Century: A Systemic Overview of Paradigm Shifts in Management. International Review of Management and Marketing, 7(3), 134-137

[7] M. Natu, S. Sethi and L. Lloyd, "Efficient probe selection algorithms for fault diagnosis", Springer, Telecommun Syst, 2008.

[8] Y. Yu, Q. Liu and L. Tan, "A graph-based proactive fault identification approach in computer networks", Elsevier, Computer Communications, 2005..

[9] 3) Çiftçioğlu, S., Almasifard, M. (2015), The response of consumption to alternative measures of financial development and real interest rate in a sample of central and east European countries. Journal of Economics, 3(2), 1-6.

[10] M.H. Lee and Y.H. Choi, "Fault detection of wireless sensor networks", Elsevier, Computer Communications, 2008.

[11] Patrick P. C. Lee, V. Misra and D. Rubenstein, "Toward Optimal Network Fault Correction via End-to-End Inference", 26th IEEE International Conference on Computer Communications, 2007.

[12] Mohamed and O. Basir, "A New Probing Scheme for Fault Detection and Identification", IEEE International Conference on Electro/Information Technology, 2009.

[13] Patrick P. C. Lee, V. Misra and D. Rubenstein, "Toward Optimal Network Fault Correction via End-to-End Inference", 26th IEEE International Conference on Computer Communications, 2007.

[13] M. Almasifard, "An econometric analysis of financial development's effects on the share of final consumption expenditure in gross domestic product", Eastern Mediterranean University, June 2013.

[14] A. Foroush Bastani, D. Damircheli An adaptive algorithm for solving stochastic multi-point boundary value problems, Numerical Algorithms 74 (4), 1119-1143, 2016.

[145] A. Foroush Bastani, Z. Ahmadi, D. Damircheli, A Radial basis collocation method for pricing American options under regime-switching jump-diffusion models, Appl. Numer. Math. 65, 79–90, 2013.