Unknown Virus Detection Method
Amalgamation Genetic Algorithm into Ant Colony Algorithm

Liu Guozhu
College of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, China
E-mail: LGZ_0228@163.com

Shang Yanjun
College of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, China
E-mail: qdsyj1983@163.com

Abstract—In order to detect new virus, a method based on biological immune principle is presented in this paper. On this basis, the clustering method combining genetic algorithm and ant colony algorithm is adopted to detect computer virus. Experiments results indicate that this method has strong adaptability, better intelligence, and higher degree of automation. Therefore, the method can be used to detect an unknown virus and possesses application value.

Index Terms—biological immune; genetic algorithm; ant colony algorithm; clustering

I. INTRODUCTION

With the popularity of computer and the expansion of the network, the computer systems attacked by virus are more and more. The phenomena can be seen from: the reproducing rate of the virus becomes faster and faster, polymorphism of viruses is developed rapidly, the spread of the network virus becomes flooding, etc. At present, the traditional methods for computer virus detection are mainly the characteristic code method, checksum method, behavior surveillance method. The lack of above methods is that they can not be used to detect new and unknown viruses completely, have high false rate and low capability. The situation often can be seen that a new virus has diffused and infected a lot of systems, the antivirus software is published late. It shows that traditional methods are obviously not qualified the new form of anti-virus technology development. Therefore, it is urgent that new virus detection technology to conquer the shortcoming of traditional methods needs to research and develop.

Biological immune system is a highly distributed, parallel and self adaptive system, and has many excellent characteristics in keeping system working in order and avoiding virus infection[1]. Naturally, it inspires us to use the idea of biological immune system to solve the problem of computer virus detection. We can use the principle, structure and algorithm of the biological immune system for the virus detection in the computer system.

Based on the principle of biological immunity, a new method to construct self and non-self is presented in this paper, which can be used to optimize and cluster system process through fusing genetic algorithm into ant colony algorithm, and monitor the possible virus invasion through the surveillance to process abnormality. Any specific virus information doesn’t need to be known in advance. So, unknown virus can be detected, and the experiment for the virus testing is carried out.

II. THE BASIC PRINCIPLE OF BIOLOGICAL IMMUNE SYSTEM

A. the Characteristics of Biological Immune

The living beings protect themselves from the harm through their own immune system. When the external virus invades the body, the biological immune system can identify the “own” and “non-own” antigen through tissue lymph cells, and remove the foreign bodies [2].

B. the definition of self

The core question of immune computation is how to define self and non-self. The characteristics of the protected system must be reflected by self integrally, namely the relatively stable characteristics of system, while non-self is bound to reflect the abnormal characteristics of the system (for infection) [3].

III. THE THEORY OF CLUSTERING ANALYSIS

The clustering is that whole data set is divided into different group. The principal is that make the difference between groups as big as possible, and make the difference within group as small as possible[4]. The typical clustering methods include division method (for example k-means and PAM), hierarchical clustering method (for example AGNES and DIANA), and density

This work has been funded by education department of Shan Dong Province P.R. China, corresponding author is Liu Guozhu.
clustering method (for example DBSCAN) etc.

The description of clustering method is that if sample set is \( X=\{X_i, i=1,2,\cdots,n\} \), here \( X_i \) is a d dimensions vector. So, the clustering problem is to find a division \( C=\{C_1, C_2, \cdots, C_m\} \). The division \( C \) is satisfied with

\[
C_i \neq \emptyset, \quad C_i \cap C_j = \emptyset, \quad i, j = 1,2,\cdots,m, \quad i \neq j
\]

and make the sum of difference within group smallest

\[
J = \sum_{k=1}^{m} \sum_{X_i \in C_k} D_{is}(X_i, Z_k)
\]

i.e., here \( Z_k \) is clustering center of group \( C_k \), \( k=1,2,\cdots,m \), and \( D_{is}(X_i, Z_k) \) is the distance from sample \( X_i \) to clustering center of group \( C_k \), i.e. \( D_{is}(X_i, Z_k) = \| X_i-Z_k \| \). The clustering objective function \( J \) is a total distance sum from every sample to its clustering center. The clustering center of every group is \( Z_k \).

\[
Z_k = \frac{1}{|C_k|} \sum_{X_i \in C_k} X_i
\]

\( Z_k \) is expressed with \( \sum_{X_i \in C_k} X_i \), here \( |C_k| \) is the amount of sample in group \( C_k \).

IV. THE PRINCIPLE OF GENETIC ALGORITHM AND ANT COLONY ALGORITHM

A. description of the Genetic Algorithm

Genetic Algorithm can be defined as a set SGA of 8 components,

\[
SGA=\{C, E, P_0, M, \Phi, \Gamma, \psi, T\}
\]

Here, \( C \) is a individual coding method, \( E \) is a individual fitness evaluating function, \( P_0 \) is an initial colony, \( \Phi \) is a selection arithmetic operator, \( \Gamma \) is a Cross-cutting arithmetic operator, \( \psi \) is a Mutation operator, \( T \) is a Genetic Algorithm end condition.

The algorithm is described as below.

Procedure SGA

Begin
Initialize \( P(0) \);
\( t=0 \);
while \((t<<T) \) do
for \( i=1 \) to \( M \) do
Evaluate fitness of \( P(t) \);
End for
for \( i=1 \) to \( M \) do
Select operation to \( P(t) \);
End for
for \( i=1 \) to \( M/2 \) do
Crossover operation to \( P(t) \);
End for
for \( i=1 \) to \( M \) do
Mutation operation to \( P(t) \);
End for
\( t=t+1 \);
end while
end

B. the principle of Ant Colony Algorithm

Ant Colony can find the shortest path from ant nest to the food resource when ant colony is in process of looking for food. Through observing and research, ant can release a special kind of secretion (information), when ants look for foods. The secretion can be used for feeling to other ants, and direct other ants move along high density direction of the material\([8]\). The ants shift probability is determined by the amount of the secretion in the path. Some path is more short, the ants are more along the path. So, the secretion is more on the path, and the probability is big to find food along this path. The probability to select by other ants is bigger. According to the rule of the path is shorter than other along the thicker information element path, the optimized path is determined.

Ants colony algorithm is described as below. \( m \) is ants amount in ant colony, \( d_{ij}(i,j=1,2,\cdots,n) \) is the distance from food resource \( i \) to \( j \), \( b(t) \) is ants amount in food resource \( i \) at the moment of \( t \).

\[
m = \sum_{i=1}^{n} b_i(t) \cdot r_y(t)
\]

is defined as information amount in the path at the moment of \( t \). At the initial moment \( \cdot C(t) \) \( (C \) is a constant \), i.e., information amount is equivalent on various paths\([9]\). When ant \( k \) is in the process of moving, the move direction is determined by the information amount in the path. \( p^k_{ij}(t) \) is the ant \( k \) moving probability from location \( i \) to location \( j \) at the moment of \( k \). \( p^k_{ij}(t) \) is expressed as below.

\[
p^k_{ij}(t) = \max \left\{ \begin{array}{ll}
\frac{\tau^a_{ij}(t) \eta^\beta_{ij}(t)}{\sum_{k \in allowed_k} \tau^a_{ij} \eta^\beta_{ij}(t)}, & j \in allowed_k \\
0, & \text{other}
\end{array} \right.
\]

Here, \( allowed_k = \{0,1,\cdots,n-1\} \) is ant \( k \) next position allowed to select. Through \( n \) moment, ant finish one loop. Information amount on every path is adjusted by below expression.

\[
\tau_{ij}(t+n) = \tau_{ij}(t) + \Delta \tau_{ij}
\]

Here, \( \Delta \tau_{ij} \) is the information amount leaved on the path by ant \( k \) from place \( i \) to \( j \) in this loop.

\[
\Delta \tau_{ij} = \sum_{k=1}^{n} \tau_{ij}^k
\]

\( \Delta \tau_{ij} \) is the increment of information amount on the path from place \( i \) to \( j \) in the loop.

\[
\Delta \tau_{ij}^k = \left\{ \begin{array}{ll}
\frac{Q \cdot \text{if the ant } k \text{ in this loop}}{L_k}, & \text{if the ant } k \text{ in this loop} \\
0, & \text{other}
\end{array} \right.
\]

Here, \( Q \) is a constant, \( L_k \) is the length of ant \( k \) in this
At the initial moment, \( t_i(0) = c, \Delta \tau_i = 0 \).

\( \alpha \) and \( \beta \) are the difference effects of accumulated information and enlighten factor in the ant moving process, when ant select path. \( \eta_{ij} \) is the expectation degree, when ant from place \( i \) to \( j \). The value of \( \eta_{ij} \) can be determined by some enlighten algorithm.

V. THE AMALGAMATION OF GENETIC ALGORITHM AND ANT COLONY CLUSTERING ALGORITHM

Genetic algorithm, formed by simulating biological process of genetics and evolution in the natural environment, is an adaptive, globally optimized and probabilistic search algorithm [5]. It was firstly proposed by Holland, a professor of Michigan University in the United States, originated in the 60s of natural and artificial adaptive system research.

Ant colony can always find a shortest path from ant nest to food source when foraging. After a large number of observations and studies, we find that ants can release an ant-specific secretion named pheromones on the way to finding food. A certain range of other ants can feel the pheromones and move towards high intensity of pheromones [8]. According to the principle of “The more concentrated pheromone, the shorter of the line”, they are able to choose the best route.

The basic idea of amalgamation of Genetic algorithm and Ant colony clustering algorithm is that: based on the rapidly global search capability of Genetic Algorithm and the positive and negative feedback convergence mechanism of Ant Colony Algorithm. At first, we use the process of genetic algorithm to generate the initial cluster center of the data object, and then use the positive and negative feedback convergence mechanism of ant colony algorithm to perfect the clustering structure. The two algorithms can complement each other [11]. Figure 1 shows the speed-time curve for Genetic Algorithm and Ant Colony Algorithm. At the initial stage of search from \( t_0 \) to \( t_a \), the convergence speed of Genetic Algorithm is higher, but decreases obviously after \( t_a \). Because of the data and its own random movement, the Ant Colony Algorithm has a slow speed at the initial stage of search, and then its speed increases evidently after the movement continues for a certain time.

### VI. CLUSTERING PROCESS OF VIRUS DETECTION

To divide the data into different groups, make the gap between groups as large as possible, and the difference between the data in the same group as small as possible, this process is clustering [4]. Several clustering methods typically include k-means and PAM, hierarchical clustering method AGNS and DIANA, density clustering method DBSCAN, and so on.

#### A. Process Description

The major processes of virus detection by clustering analysis include data collection, data standardization, clustering analysis and tagging instance data. Virus detection process is described as follows:

1. **data collection** (the collection of self attributes)
   First we randomly collect \( m \) processes (recorded as sample \( G \)), each of which has \( k \) types of attributes (the relatively stable process attributes of system). For the attribute vector \( P_i = (p_{i1}, p_{i2}, \ldots, p_{ik}) \) (\( x = 1, 2, \ldots, n \)), each attribute value will be converted into the two bits of 0 or 1, string \( s_i = \{00, 01, 10, 11\} \). The transformation function \( g \) is given by formula (1), and the min and the max are the lower and upper limit of the attribute of type I in normal circumstance respectively [12].

   \[
   g(p_i) = \begin{cases} 
   00, & p_i < \min \\
   01, & \min < p_i \leq (\max + \min)/2 \\
   10, & (\max + \min)/2 < p_i \leq \max \\
   11, & p_i > \max 
   \end{cases} \quad (1)
   \]

   (i = 1, 2, \ldots, k)

   Here, \( g(p_i) \) is the instance data of process attributes, and \( P' = \{g(p_{i1}), g(p_{i2}), \ldots, g(p_{ik})\} \).  

2. **standardization of instance data**
   In order to eliminate the impact of similarity caused by the different measurements, we should standardize the measured value of instance data before using clustering algorithm. The method is as follows:
   If there are \( n \) samples, each sample has \( m \) data, then each variable can be recorded as \( X_{ij}, i = 1, 2, \ldots, n \); \( j = 1, 2, \ldots, m \) [12].

   After standardization, variable becomes \( X_{ij}', \) then
\[ X_{i,j} = \frac{X_{i,j} - X_j}{S_j} \quad (i = 1, 2, ... n) \\
\]

\[ \overline{X_j} = \frac{1}{n} \sum_{r=1}^{n} X_{i,j} \quad (j = 1, 2, ... m) \\
\]

\[ S_j = \sqrt{\frac{1}{n-1} \sum_{r=1}^{n} (X_{i,j} - \overline{X_j})^2} \]

Here, \( \overline{X_j} \) is the average for the \( j \)th variable, \( S_j \) is the standard deviation for the \( j \)th variable. \( X_{i,j} \) is the \( i \)th element of the \( j \)th variable.

3) Using genetic algorithm to initialize the cluster set

Step 1: Description of the problem

Set a objective function:

\[ \min J = \sum_{r=1}^{P} \sum_{i=1}^{m_r} \left\| X^{(r)}_i - C_r \right\|^2 \]

Here, the clustering center \( C_r \) is:

\[ C_r = \frac{1}{m_r} \sum_{i=1}^{m_r} X^{(r)}_i \quad (r = 1, 2, ..., P) \]

\[ \sum_{r=1}^{P} m_r = N \]

Step 2: Structure of chromosome

Let \( \overline{Y} = (S_1, S_2, ..., S_L) \) be the structure of chromosome, \( \overline{Y} \) is a \( 1 \times L \)-dimensional row vector, and \( S_i \) (\( i \leq L \)) is the No. \( i \) gene. Let \( N \) be the sample size, and the chromosome requires the conditions as follows:

(I) \( L = N \);

(II) \( S_i = (1, 2, ..., P) \), \( i = 1, 2, ..., L \); (6)

(III) \( \sum_{i=1}^{N} d(r) = P \) here \( d(r) = \begin{cases} 1 & \text{if } r \in \{S_1, ..., S_L\} \\ 0 & \text{if } r \notin \{S_1, ..., S_L\} \end{cases} \) (7)

Step 3: The fitness function and genetic operator operation

Fitness function \( F \) is defined as \( F = M / J \), in which \( M \) is a constant, and \( J \) is defined as equation (2). This individual with small value will have high fitness value correspondingly.

The processes of genetic operators operation include selection, crossover, mutation, and so on.

a) Selection rules: In accordance with the fitness function above, we use the roulette strategy to choose[15]. The best chromosome of each generation will be chosen firstly in the next generation, and the number of the same chromosome of each generation chosen in the next generation is no more than 2.

b) Cross-cutting rules: Cross rules adopts the method of digital crossover control[14]. But there may appear the abnormal conditions that the chromosomes after cross-cutting are not satisfied with the condition (7). So, we set the maximum number of attempts allowed. If the number exceeds, we should abandon this matching chromosomes and re-matching for the cross.

c) Mutation rules: the uniform mutation method will be employed[13]. We assume that there is a individual \( X \), let \( X = x_1x_2...x_n \), if \( x_k \) is the mutation point, which value range is \( [U_{min}^k, U_{max}^k]\). After uniform mutation operation to individual \( X \) at the point \( x_k \), we will get a new individual \( X' = x_1x_2...x_k...x_n \), in which the new gene value of the mutation point is:

\[ x_k = x_k + \tau \cdot (U_{max}^k - U_{min}^k) \]

here, \( \tau \) is a random number in the range of \( [0,1] \) which in line with the uniform probability distribution.

Step 4: Obtaining the initial clustering set through using genetic algorithm.

4) Using the ant colony clustering algorithm to cluster and optimize clustering set which has been initialized by the genetic algorithm.

Let \( X = \{X_i|X_i=(x_{i1}, x_{i2}, ..., x_{im})\} \), \( i=1,2,..N \} \) be a data instance set which will be cluster analyze and the set has been clustered by the genetic algorithm. The algorithm process based on the ant colony optimization and clustering is as follows:

Step 1: Assigning values to variables: \( \alpha=1, \beta=1, \tau_0(0) = 0 \), that is, at the initial moment the amounts of information of each path are equivalent to 0. For variables \( \epsilon, \tau_0, \eta_0, \eta_0, P_0 \), we set different values respectively to do clustering experiments. Here, \( m \) is a number of instance attributes, \( r \) is the cluster radius, \( \epsilon_0 \) is the lower error limit, and \( \eta_0 \) is the lower probability limit of the merging cluster[10].

Step 2: Calculating the distance according to equation (8).

\[ d_{ij} = \left\| P(X_i - X_j) \right\|^2 = \sum_{k=1}^{m} p(x_k^i - x_k^j)^2 \] (8)

Here, \( d_{ij} \) is the weighted Euclidean distance from \( X_i \) to \( X_j \), \( p \) is the weighted factor, which can be assigned according to the difference of the contribution in the cluster. We calculate the amount of information of path \( i-j \) at the moment \( t \).

\[ t_{ij}(t) = \begin{cases} 1, d_{ij} \leq r, \\ 0, d_{ij} > r. \end{cases} \] (9)

Step 3: Calculating the probability after \( X_i \) integrated into cluster \( C_j \) according to the formula (10).

\[ p_{ij}(t) = \frac{\tau_0^a (t) \eta_0^b (t)}{\sum_{k=allowed} \tau_0^a (t) \eta_0^b (t)} \] (10)

If \( p_{ij}(t) \geq P_0 \), \( x_i \) would be integrated into \( C_j \). Let \( C_j \)
Step 4: calculating the center of the cluster according to the equation (11), here, $X_k \in C_j$.

$$C_j = \frac{1}{s} \sum_{k=1}^{s} X_k$$

Step 5: calculating the deviation error of the $j$th clustering by equation (12). $c_j$ is the $i$th component of the $j$th cluster center; and $x_{kj}$ is the $i$th component belonged to instance $X_k$ of cluster $c_j$.

$$D_j = \sum_{k=1}^{s} \sqrt{\sum_{i=1}^{m} (x_{ki} - c_{ji})^2}$$

Step 6: calculating the overall error after clustering according to equation (13).

$$e = \sum_{k=1}^{M} D_k$$

If $e \geq e_0$, switch to step 2; otherwise stop the process of ant colony optimization. As the size of normal instance is far greater than the number of abnormal instance under normal conditions, we can assume that the number of category and its subcategory which show normal behaviors is much bigger than then size of the category and its subcategory reflecting aggressive behaviors. After optimization and clustering of the ant colony are completed, all clusters are sorted by the size of instances each cluster included. Setting the threshold value as $B$, if the cluster’s size greater than $B$, we will see the cluster’s behavior as normal behavior, otherwise as abnormal behavior. Finally the corresponding results of virus detection will be got.

B the diagram description of clustering algorithm fused

The block diagram of clustering algorithm fusion is shown by figure 2:

**Figure 2. the block diagram of clustering algorithm fusion**

VII. EXPERIMENTAL RESULTS

The unknown virus detection experiments have been done on the PC-windows NT platform, and five protected programs have been chosen: Rundll32, lsass, ping, csrss, winlogon. There are 12 kinds of process attribute: the execution time of user mode, the execution time of the system mode, authority credentials, the presence of the page, the cumulative number of changed pages, the number of pages in virtual memory, the number of body pages, the number of data pages, the number of pages in stack, the number of documents open in the process, the number of TCP and UDP port open in the process and its sub-process.

In this experiment, we used a file infected virus as non-self, and let 48 family viruses(containing a total of 148 deformed virus) infect system files. Firstly, we use the virus detection module to identify the virus according to the algorithm mentioned above. Then we use the virus detection program of traditional feature code engine, the virus detection program of checksum engine and the virus detection program of behavior monitoring engine to detect viruses respectively. The testing effects of four ways are shown in figure 3.
What's more, the testing results of the experiments also show that this method has good effect on unknown virus detection, and will have great application values on the future study of virus detection technology. The method of virus detection presented in this paper has good detection capability under the same conditions of unknown virus, the detection rate of the feature code is the lowest. This is because the feature code engine need the virus’s pattern files when detects the virus. If there are no pre-defined virus pattern files, its detection rate would be lower, and false alarm rate would be very higher. Checksum method would check the consistency of the checksum kept originally and the checksum calculated by checking the present content of documents regularly or before using the files, which may mistake the legal access to documents for abnormal behavior. So the false alarm rate of checksum method is very high. The behavior monitoring method has summed up some common behaviors of the virus after the observation and study to the virus for many years. When the program is running, it monitors the behaviors, and alarms if finding the virus behavior. This method plays a role only if finding the virus characteristics in common, that is to say, the accumulating experience is to be needed. When new virus appears, this “experience” would not work well, and the effect of the unknown virus detection is greatly reduced. Figure 3 has reflected that the method presented in this paper has good detection capability.

VIII. CONCLUSION

The method of virus detection presented in this paper emphasizes the definition of self from the dynamic behavior attributes of the process, and the clustering of the process attributes. This approach has greater advantage than the traditional virus detection technology. What's more, the testing results of the experiments also show that this method has good effect on unknown virus detection, and will have great application values on the future study of virus detection technology.

REFERENCES


Liu Guozhu was born at Qingdao, Shandong province, China in 1965.2. He received bachelor degree in applied math from East China University of Science and Technology (QUST) in 1986.7. He received master degree in computer application from Ocean University of China in 1995.7. His major fields of study include computer network security and video compress algorithm etc.

He joined Qingdao University of Science and Technology in 1986.7. He was an associate professor from 2000.10 in QUST. He have published more than 30 papers.

Shang Yanjun was born at Qingdao, ShanDong province, China in 1985. He is a master in QUST. His major field of study is virus process