Mathematical modeling method based on genetic algorithm and its applications

L V Stepanov, A S Koltsov, A V Parinov and A S Dubrovin
ET Department, Voronezh Institute Federal Penitentiary Service of Russia, 1a Irkutskaya Street, Voronezh, RU

E-mail: noev87@mail.ru

Abstract. In many spheres of human activity are having challenges and amounted to an identification of the most rational solution, taking into account the prevailing conditions and limitations, that is, the optimization problem. Such areas could include not only socio-economic activities, but many of the engineering industry. To date, invited and actively used a large number of mathematical methods of solution of optimization tasks. The success of the solution of practical tasks depends on the quality and efficiency of applied mathematical methods and models. A very important consequence of mathematical modeling is to implement a synthesized model in the form of algorithms and programs for electronic computing machines. Describes a way of formalizing the various kinds of information security threats and vulnerabilities of the information system of enterprises and institutions. Because of the complexity of ensuring the protection of the information of any organized system should apply new concepts and solutions in the field of information security. One such approach is a method of genetic algorithm. For enterprises of all spheres of activity is an actual question of integrated evaluation of level of security of information systems, taking into account the quantitative and qualitative factors that characterize the components of information security.

1. Introduction
One of the methods commonly used to construct mathematical models include linear, dynamic methods, as well as integer programming, theory of graphs, and others.
However, on the one hand, the ‘classic’ technique may demonstrate a lack of effectiveness in practical tasks, distinguished by a large number of options that require accounting. In these circumstances, the model significantly expand, and variability of these parameters can lead to the need for periodic adjustment of mathematical models. Putting them in the form of algorithms and programs then loses meaning. On the other hand, mathematical methods themselves could significantly complicate subsequent practical implementation results. These methods can be used, for example, include those based on the application of differential calculation. The solution of these equations are very difficult to implement with ordinary tools in application software.

Among the mathematical methods and models that appeared relatively recently can be attributed applied methods of decision theory, game theory, methods, neural networks, genetic algorithms, and some others.
Among these approaches are also not all may have subsequent practical implementation in the form of algorithms and programs for electronic computing machines. [6]. for example, the theory of differential games contains games that are applicable in cases where the system parameters can be represented as functions. This kind of differential calculation problems inherits games. In addition, the terms variability structure interaction of practical implementation models games becomes a question. And in the presence of uncertainty arises a large number of limitations to obtain the solution. Methods of neural technologies in these circumstances is also not effective due to the need to design and subsequent training new neural network at change, for example, the number of input parameters.

In the above review of the particular situation takes genetic algorithm method [2, 5]. There are also names: evolutionary programs, evolutionary strategies and genetic programming. This algorithm uses the idea of natural selection 5th among living organisms in nature, therefore it is called genetic. The special status of the genetic algorithm is that it steps may be implemented in the form of algorithms and programs. Principles of similarity naturally biological processes and their mathematical formalization need separate considering, however, that not all biological mechanisms can be described analytically.

2. The main stages of the genetic algorithm method
Genetic algorithm method is a generalized description of the process in biological systems [2] figure 1.

![Figure 1. Diagram of the genetic algorithm.](image)

As the dignity of the genetic algorithm secrete a number of features in comparison with the classical methods used to solve optimization problems. First, the genetic algorithm handles encoded values of the parameters of the task. Secondly, finding a solution is not from any starting point. It iteratively occurs based on a certain set of values that describe the problem, thus avoiding stops searching the locations of local extrema. Third, use a function, rather than any indirect data. This attests to the stability of the genetic algorithm and positively differentiates it from other technology solutions to optimization problems [2]. In addition, it is necessary to note the presence of the genetic algorithm of a significant number of parameters that can be set up under the terms of a specific task. These parameters are present at all stages of the algorithm.

The first is initializing the chromosomes, that is the formation of the initial population. The formalization of this process is to define the initial set of vectors that characterize the State of the
system being modeled. There are different forms of presentation of population (initial set) and its chromosomes (vectors).

The second stage is assessing the fitness of the chromosomes in the population. This evaluation is the calculation of the value of some characteristic function for each source vector set. In biological systems, the higher the fitness value, the higher the quality of the chromosome. The function should reflect this principle.

In the implementation phase of the breeding also used the fitness function value. The essence of this phase is to choose pairs of chromosomes (vectors) of the population (initial set), which will participate in the creation of the descendants (the next population or new generation). In biological systems, this process is random (as installed) by. The result of the selection is considered to be set (the parent pool), which is equal to the number of vectors of the instantiated many, but its elements are divided into pairs. Applying genetic operators to chromosomes is the core of the genetic algorithm. This will result in a new population of descendants. At this stage, crossing statement executes and the operator mutation, which makes random modification as the result of crossbreeding. It is important to note that the mutation in biological systems is not negative values, and a mechanism for implementing diversity. In relation to mathematical realization of this phase, the mutation is to accidental modification of any vector (one or more of its members) before or after mating. The essence of cross-breeding is to identify one or more-check Division of vectors formed pairs and combining parts of different vectors in a couple new vectors.

The fifth stage is the formation of a new genetic algorithm population according to the principle of natural selection. A new population (many vectors) is formed from individuals (vectors) with the best fitness (the value of the characteristic function), i.e. the source population is decreasing and this set becomes the outcome for a new iteration genetic algorithm.

Next is determined by the possibility to stop the genetic algorithm. Biological systems, this occurs if the populations detected chromosome with the highest relevance, i.e. a vector with a maximum value of characteristic functions. In real practical task, this condition may be different.

If genetic algorithm cannot be stopped, all stages starting from the evaluation of fitness. If genetic algorithm can be stopped, the best solution is considered a chromosome (vector) with the greatest or least value fitness function. This vector represents the optimal solution to a problem.

3. Influence of phase crossing on genetic algorithm

As noted above, the central phase of the genetic algorithm is crossing. It is held with a view to obtaining a new population of chromosomes. The functionality of this phase can be carried out analogy with the process of mutual physical dissolution Wednesday.

Factor will definitely guarantee the emergence of a new unique features is an arbitrary change of one or more parameters crossing vectors, i.e. the process of mutation. However, in applied areas of need for intentional or accidental changes of parameters characterizing the State of the system or process. Moreover, it can be interpreted as a significant distortion of the source data for the task.

The process of interbreeding is that for each pair of chromosomes (vectors) is determined by the position of a gene (the value in the vector) which determines the point of crossing. It is the same for both s crossing chromosomes (vectors). In biological systems, this process occurs randomly for each of the chromosomes, but the position is always the same (except for the case of deviations). At this point the gap of chromosomes occurs (vectors) and the subsequent formation of new two pairs, but made up of parts of different vectors. Length of educated children must be even length 'ancestors'. When multipoint determines the number of crossing points crossing more than one. Sometimes consider even crossing [2]. When this option is selected, the standard that sets the values of the vector must be inherited from the first, and what of the second vector.

There are two components:
1. Select how to define the position of the point (or points) of the crossing;
2. Analysis of the impact of the number of points on a genetic algorithm.

Determination of mating position with a view to ensuring the likes of biological systems should be in a random way. Any management can reduce the accuracy of the results and may consider as deviation from the biological genetic algorithm. The increase in the number of crossing points will obviously lead to ascending result accuracy. This effect can be explained by the fact that there is a growing degree, i.e. the mutual penetration of chromosomes (figure 2). ‘Child will best reflect the original characteristics of the’ ancestors’.

Original vector 1: [001100111010]
Original vector 2: [101011011011]
The crossing point: 4 and 6

Vector 1: [0011 00 111010]
Vector 2: [1010 11 011011]

4 6

The destination vector 1: [001111111010]
The destination vector 2: [101000011011]

Figure 2. Two-point crossing. Binary version.

Thus the choice of multi-point mating should be seen as preferred.

In these circumstances, should decrease the number of iterations and consequently the load on the computer system in the case of automated version of realization of genetic algorithm [6]. Especially when a large number of crossing points and a large number of chromosomes of the source population. To study the problem machine was carried out the experiment. The original lot consisted of 200 vectors. The value of each generated randomly in the range from zero to fifty. Used as a breakpoint condition approximate equality amounts chromium-skreshhivaemyh som items. The study showed that the change in the number of crossing points from one to fifty does not significantly affect the number of iterations until convergence algorithm. A series of such experiments confirmed this result. However, the conclusion may be true for these parameters by genetic algorithm.

4. Practical aspects of using genetic algorithm
In spite of the dignity of the genetic algorithm in the analysis stages, we can conclude that the algorithm parameters must be justified (figure 3). So will highlight aspects that need defining according to the stages of the genetic algorithm in the application:

1. initialize a population of chromosomes—you must consider the effects of various forms of presentation of population (initial set) and its chromosomes (vectors):
   - vectors can be represented by a binary representation of the values of the sequences. binary vectors can be considered typical. The best analogy is found with genotype in biological systems. If the original values of the parameters characterizing the system specified in various
scales, before translation to binary form should normalize the values corresponding to the same scale [3, 7].

-Vectors can be presented as valid values. In this case also reveals the analogy with biological systems, but with the phenotype. The need to normalize the values and the same scale also saved [3, 7].

-vectors can be represented by Fuzzy (linguistic) values. Dan Mr. case, you can use approaches to formalisation of linguistic structures and transfer them into numeric form [4, 7]. These techniques enable you to solve the problem heterogeneity of vectors of values of the initial set.

It should be borne in mind that the vectors of initial population may not have a different number of elements. This assumption can be substantiated by the fact that the original lot describes the State of the same system or process. If the task of optimization examines the interaction of two or more parties, the original population can be represented by the appropriate number of subsets of vectors.

| Initializing the population of chromosomes | Population reporting forms (initial set) and its chromosomes (vectors) |
|-------------------------------------------|---------------------------------------------------------------------|
| Evaluation of adaptability of chromosomes  | View fitness function                                               |
| Selection of chromosome                   | Method of forming pairs of chromosomes (vectors)                   |
| The application of genetic operators      | -The desirability of implementing the mutation.                    |
|                                           | -option: single point of crossing, multi-drop or more.             |
| The formation of a new population         | Criterion and method of selection of the chromosomes (vectors) in new population (many vectors). |
| Stop algorithm                            | A new population with a pair of chromosomes of some importance of fitness, the expiration of a certain period of time, the execution of the specified number of iterations |
| Choosing the 'best' chromosome            | Solution selection criteria                                        |

**Figure 3.** Parameters of genetic algorithm.

2. evaluation of the fitness of the chromosomes-it must be borne in mind that the type of the function affects the evaluation of the fitness of the chromosomes (vector). However, an important advantage of the genetic algorithm is that one and also feature is used to calculate the characteristic values of all vectors of the initial set. Consequently, the impact on the effectiveness of a genetic algorithm will not. It is important that you are adequately function membrane aggregation of the values of the vector.
3. selection of chromosomes—requires the choice of methods of forming pairs. In biological system, this process can be coincidental. To implement a random way to apply method Roulette [2]. He is considered to have the best implements the semblance of evolutionary principles. The value of sector notional roulette selected proportional to the value of a function, as the individual with the best fitness has great chances of children. In other systems, method of forming pairs significantly depends on the problem.

4. genetic operators to select modification principle vectors. To better meet this operation should be carried out randomly. However, practical problems in the first place it is important to take a decision of principle on realization of mutation, because in fact we are talking about the distortion of the original data. The prerequisites and consequences of crossing option analysed above.

5. formation of new population will have a major impact on genetic algorithm, because at the current stage of decline many vectors. You want to precisely formulate criteria and method of forming a new population. In fact, when the realization of natural selection can be lost critical information and it would have a negative impact on solving the optimization problem. Most often the condition this task lets you decide what value the function vectors fitness should be deleted. For example, if the task of maximizing, the vector with a minimum value of should be deleted. When solving the task of minimizing. Combined option is possible.

6. Stop condition genetic algorithm also depends on specific conditions. The algorithm can be stopped upon receipt of a new population, containing a pair of chrome-catfish, certain fitness function value after a certain time or execution after the specified number of iterations. Any of these options is not contrary to the natural processes of nature.

7. Choosing the optimal solution task—from the standpoint of biological systems chromosome takes precedence with the greatest relevance.

as can be seen from the analysis at each stage has one or more parameters that must be specified when the task of mathematical modeling using genetic algorithm. The mere presence of these parameters also suitable genetic algorithm differs from other methods.

5. Formalization of vulnerabilities and threats to information security enterprise or institution
The threat is a consequence of enterprise information system vulnerabilities. While the composition and structure of the threats and vulnerabilities may not be the same. For the earlier example about malicious software vulnerability in specifically information system may occur due to the lack of technical means of protection required quality.

We assume that \( H \) and \( A \) many threats and many information security vulnerabilities:

\[
H = \{H_j\}, j = 1, m, \quad A = \{A_i\}, i = 1, n
\]

where \( H_j \) - j-th the threat of information security; \( A_i \) - i-th information security vulnerability.

While, \( H_j \) define as:

\[
H_j = \{h_j^y\}, y = 1, w_j,
\]

where \( h_j^y \) - th feature j-th information security threats, that is, action or event capable of lower level of information security; \( w_j \) - the number of characteristics of j-th threat.

Due to the fact that the vulnerabilities and threats can be described not in numeric form, using linguistic constructions, it is proposed to divide their entity for quantitative and qualitative.

The difference between quantitative and qualitative parameters. For example, failure of equipment may well be evaluated quantitatively in terms of value, but the corruption, and, a fortiori, lowering its confidentiality may be described qualitatively more often.
For the transition from the qualitative expression characteristics of threats and vulnerabilities to their numeric expression using the theory of decision-making, as well as the theory of fuzzy sets. The individual components are adapted to the task of ensuring information security [1].

For each quality you can specify many characteristics of linguistic structures, accurately and completely describes the values that can be characteristic.

To switch to numeric values you want to correspond to the characteristics of each linguistic constructs some weighted value. It must take into account the correlation between linguistic constructions.

Generate a square matrix [1] available pair-wise comparisons between the linguistic structures:

\[
SR = \begin{bmatrix}
    s_{r_1} & ... & s_{r_b} & ... & s_{r_r} \\
    ... & ... & ... & ... & ... \\
    s_{r_b} & ... & s_{r_b} & ... & s_{r_r} \\
    ... & ... & ... & ... & ... \\
    s_{r_r} & ... & s_{r_b} & ... & s_{r_r}
\end{bmatrix}
\]

where \[ sr = \begin{cases} 
    sh, l_b > l_b, \\
    \frac{1}{sh}, l_b < l_b, \\
    1, l_b = l_b.
\end{cases} \]

In this expression, a symbol of «» indicates that one of linguistic structures dominates the other.

On the basis of the SR it is possible to calculate the weight for each qualitative values, for example, by summing the elements in rows or columns of the matrix

\[
Sum_b = \sum_{b=1}^{r} sr, b = \bar{1}, r,
\]

where \( b \) and \( b' \) - the row and column indexes of the matrix \( SR \).

Normalize values:

\[
o_b = \frac{Sum_b}{Sum_{oben}}.
\]

For future reference, you also need to normalize quantitative values of initial sets of threats and vulnerabilities. The result will be many vectors threats and many vectors A of information security vulnerabilities, represented by numeric values.

6. Getting the resulting characteristics of the enterprise information security level

Received multiple vectors of H and A systemically interconnected. As noted above, the threats are a consequence of the vulnerabilities of the information system. Therefore, for any business you can create a basic set of parameters that characterize all aspects of information security [5, 6].

Information security options, not occurring for a particular threat or vulnerability can formalize the value zero. This will allow you to get many H and A, vectors which will contain an equal number of elements. According to the number of multiple vectors themselves (H) and (A) may be different. In practice, this may be due to the fact that the information system can be implemented more measures to ensure information security than the currently identified threats. Consequently, many vulnerabilities will be less. In these circumstances, to set you want to apply a method that would 'cover all the elements of sets and (A) irrespective of the number of items in them.

This method is a method of genetic algorithm, features and modifications which are discussed in a number of works [2, 3, 4]. Principles of genetic algorithm does not require special elaboration. The algorithm allows not only the transition from genotype to phenotype, i.e. the replacement of real numeric values, but the binary code and can be implemented at the level of the phenotype, i.e. at the level of actual numeric values. This feature is defined at the initial stage of the algorithm. Because the original population are based on normalized values, threats and vulnerabilities, need to move to binary values are not.
However, the application of genetic algorithm to the sets $H$ and $A$ needed discussion [2].

The first step is to initialize the formation and formalization of the original sets $H$ and $A$.

The second stage consists in assessing the fitness of chromosomes, i.e. the calculation of the values characteristic of each vector sets $H$ and $A$. If all components of the threats and vulnerabilities are equivalent to obtain fitness function value offer use geometric mean or the arithmetic mean of the dependence of calculated based on normalized values of each chromosome characteristics:

$$
\Pi_j = \sqrt[|w|]{\prod_{y=1}^{w} h_j^y}, \quad \Pi_j = \frac{\sum_{y=1}^{w} h_j^y}{w'}
$$

(6)

The third phase is determined by the possibility to stop the genetic algorithm. As this condition in relation to the current task and the need to narrow the initial sets of threats and vulnerabilities can be a time of receipt of the last of the total vector describing the population as a whole [3, 4].

If the stop condition is not performed, is the fourth phase of selection of the chromosomes, i.e. selecting pairs of chromosomes (vectors) to be interbred. The most popular is considered the so-called roulette method [2], selecting pairs when applied has casual character.

In the next step to a pair of vectors are applied genetic operators, the essence of which is to obtain descendants based on the original chromosomes. For each pair (threat-vulnerability) is determined by one or several points of crossing. The increase in the number of points increases the complexity of the genetic algorithm, but it leads to increased accuracy of the result. This is due to the fact that this increases the degree of mutual penetration chromosomes. Descendants received better reflect the original characteristics of parental chromosomes.

The next major step in the genetic algorithm is the stage of the formation of a new population. At this stage the chromosomes that results from using the crossing operators to chromosomes of the temporary population is included in the newly formed populations, which becomes the new generation for the genetic algorithm. Shaping is done on the basis of natural selection. According to him, most likely to create descendants most individuals of relevance. In terms of biological systems, this principle is unequivocal. However, its use in other contexts can be interpreted otherwise [3, 4]. So, when evaluating information security level the most characteristic representatives of the population threats and vulnerabilities can be considered a chromosome from the top fitness for threats and vulnerabilities. It means that the selection process should build on the exclusion of chromosomes (the elements of the original sets) with a minimum value of fitness for $H$ and $A$ as representatives of the least characteristic for each population.

Cumulative value characterizing information security at enterprise or institution is considered a chromosome with a high value of fitness. This vector is a vector of task solution, describing the information security across the enterprise.

Assuming that the initial $H$ and $A$ population can have a different number of individuals (vectors) of genetic algorithm applied to them can be done in different ways:

1. application of genetic algorithm to separate initial population and separately to initial population $H$ and $A$ to determine the resulting vectors threats and vulnerabilities and then applying genetic algorithm to these vectors;

2. application of genetic algorithm to randomly selected pairs of $H$ and $A$ to obtain a result vector of information security.

7. Results and discussion

Enterprise information systems are complex and often geographically distributed structure. A large number of staff with different levels of computer training leads to a significant growth of information security vulnerabilities [5, 6]. Along with the two selected and discussed the components that determine the level of information security (threat and vulnerability), you can also highlight events (activities) $M$ to improve information security. This activity can be regarded as the set of solutions to
attract personnel assets (institutional, administrative, programmatic, technical and legal) to address these vulnerabilities. Actions can be directly are not directed at any particular component of information security. They may be indirect, but lead to minimize or eliminate the vulnerability. For example, work with the staff on 'computer literacy' can lower the risk of exposure to harmful software.

The result is the possibility of multilateral research enterprise information security, on the basis of which the original pair sets are considered three possible \((H, A, \text{ and } M)\). The results can be interpreted in different ways:

1. the source of a multitude of \(H\) and \(M\)-allow you to evaluate the effectiveness of interventions to address threats;
2. multiple \(A\) and \(M\)-assess the effectiveness of interventions to eliminate vulnerabilities information system;
3. multiple \(A\) and \(H\), and assess the degree of relationship (interdependence) threats and vulnerabilities of the information system.

The following results in need of discussion are the functioning parameters of genetic algorithm:

1. assessing fitness values can be done in different ways. If you use geometric mean or arithmetic mean of dependency and components of chromosomes not equivalent, the weighting factor could be introduced, taking into account the nonequivalence:

\[
P_i = \sqrt{\prod_{y=1}^{w} o^y h^y_j}, \quad \Pi_j = \frac{\sum_{y=1}^{w} o^y h^y_j}{w'},
\]

where \(o^y\) - weight \(y\)-th characteristics.

The values of the weighting factor is determined prior to the application of the genetic algorithm. They must be equal to 1 and throughout the work of genetic algorithm must not be adjusted.

2. breakpoint genetic algorithm and the formation of a new population at each iteration of the algorithm-classic genetic algorithm population of entire source is replaced by a population of descendants with the same number of elements. In the circumstances, the application of genetic algorithm to get only the combined vector that characterizes the information security. It is therefore advisable to provide for a reduction in populations at each iteration [4]. Thanks to the selection procedure proposed above for each iteration of the genetic algorithm, the number of new generation decreases. As a result natural breakpoint genetic algorithm can be seen as the moment when in a population will remain only two vector values of characteristics of information security.

The final result is a vector that characterizes the information security. It will be a set of normalized values that describe the parameters of information security at the enterprise or organization. Some data will correspond to the quantitative and qualitative part. Last received data can be converted from a numeric type back to a form of linguistic structures on the basis of the system previously described interval estimates.

8. Conclusion

The application of the proposed model allows you to assess the level of information security based on identified threats and vulnerabilities, as well as measures to ensure the security of information or infrastructure or enterprise information system institutions.

This level of evaluation based on objective information about threats and vulnerabilities information system and does not require a solution to the problem of finding the characteristics of integrity, accessibility and confidentiality of information. This search can be difficult, since the integrity, availability and confidentiality is difficult to formalize. In addition, this approach is adaptive to changing information security threats and vulnerabilities, both in structure and parameters.
Identification of areas for improvement in information protection system and analysis of its dynamics could very well reveal more, or perhaps less effective measures to ensure information security in the context of the current situation can be adjust or not implement.

References
[1] David G 1978 Method of paired comparison (M.: STA-statistics) p 144.
[2] Rutkovskaya D, Pilinskij M I and Rutkowski A 2006 Neural networks, genetic algorithms, and fuzzy systems (M.: Hotline-Telecom) p 452.
[3] Saati T 1993 Decision-making-analytic hierarchy process (M.: Radio Communications) p 278.
[4] Stepanov L V 2008 Perfect competition condition assessment Modeling based on genetic algorithm Problems of the theory and practice of management 8 pp 111-118.
[5] Stepanov L V 2010 Approach to formalization process formation of the market Information systems and technology 6 (62) pp 22-27.
[6] Stepanov L V 2010 Implementation models of market competition in the software Software products and systems 4 pp 144-147.
[7] Stepanov L V 2010 Interaction model of the enterprises in conditions of monopolistic competition market Automation and modern technology 1 pp 42-46.
[8] Stepanov L V, Parinov A V, Korotkikh L P and Koltsov A S 2018 Approach to estimation of level of information security at enterprise based on genetic algorithm International Conference Information Technologies in Business and Industry (Tomsk Polytechnic University, Russian Federation) p 123-134.
[9] Koltsov A S, Larkin O V and Rossikhina L V 2017 Analysis of information processes in educational Wednesday when solving mathematically incorrect tasks (Herald of the Voronezh Institute of the Federal Penal Correction Service) 1 pp 77-80.
[10] Parinov A V, Shvartskopf E A, Popova L G, Bataronov I L and Tolstykh N N 2018 Risk models of destructive content diffusion between social network communities International Journal of Pure and Applied Mathematics 15 pp 2605–2609.
[11] Parinov A V, Shvartskopf E A, Parinova L V, Razinkin K A and Belonozhkin V I 2018 Social Information Networks: Models Of Internetwork Malicious Content Diffusion International Journal of Pure and Applied 15 pp 2639–2643.
[12] Ostapenko A G, Kalashnikov D, Carpenters G and Parinov A 2015 Expected assessment and management effectiveness and survivability of attached objects in the information field: problem statement and objectives research Information and Security 2 pp 218-231.
[13] Ostapenko A G, Parinov A and Kalashnikov A Social Networks and destructive content: monograph (M.: hotline-Telecom) 3 pp 2017-276