The application of an artificial immune system for solving the identification problem in Ecology

IRINA ASTACHOVA[1]*, STANISLAV USHAKOV[1], ANDREI SELEMENEV[1], and JULIYA HITSKOVA[2]

[1]Department of Applied Mathematics and Informatics,
Voronezh State University, Voronezh 394018, RUSSIA
[2]Department of Information Systems Design in Civil Engineering,
Voronezh State Technical University, Voronezh 394006, RUSSIA
astachova@list.ru, stan@mail.ru, andreyjkee@gmail.com, prosvetovau@mail.ru

Abstact -. Ecological prognoises sets the identification task, which is - to find the capacity of pollution sources based on the available experimental data. This problem is an inverse problem, for the solution of which the method of symbolic regression is considered. Some authors solved a similar problem with the help of neural networks. In this paper the distributed artificial immune system is used as an algorithm for the problem solving. The artificial immune system (AIS) is a model that allows solving various problems of identification, its concept was borrowed from biology. The solution is sought using a distributed version of the artificial immune system, which is implemented through a network. This distributed network can operate in any heterogeneous environment, which is achieved through the use of cross-platform Python programming language. AIS demonstrates the ability to restore the original function in the problem of identification. The obtained solution for the test data is represented by the graph. This language has been chosen for the following reasons: cross-platform reason: during the creation of an application, running in a heterogeneous computing environment, the cross-platform factor is very important; the speed of development reason: Python is oriented on the increase of developer productivity and code readability; the standard library reason: includes a large number of useful functions and classes for various tasks such as networking, multithreading, process management.

Keywords − Artificial Immune Systems (AIS), symbolic regression, distributed calculations, identification problem in ecology.

1 Introduction

Symbolic regression – is a method of construction of regression models by enumeration of various arbitrary superpositions of functions in a given set. The superposition of functions is called "program" and the stochastic optimization algorithm for constructing of such superpositions is called genetic programming. However, instead of commonly used genetic algorithms the artificial immune system will be used.

The problem of finding of the optimal regression structure model will be formulated further. Thus, the input data for the problem stated is the collection of points of n-dimensional space and values of the function at a point. The problem solution - is a function, which depends on *n* variables, that best approximates the original (at fixed points). To solve the problem of symbolic regression the algorithms of genetic programming are used [17]. Genetic programming is a way to create programs with the use of genetic algorithms [18]. In this case, the program will be presented by the function in the form of a tree, which is allegorically represent a chromosome itself. Crossover operators and mutations occur over the tree of function representation [17], the adaptation function is a measure of how well the program (function) solves the necessary task.

Lets consider an artificial immune system, which will be used for solving symbolic regression problems, and lets take a look which of the basic properties of natural immune system will be useful for solving this problem. At the second step, after the creation of a simple model of the immune system, we use the fact that the immune system is inherently distributed, thus we modify the algorithm in such a way, so that it uses distributed computing that can speed it up and improve.

1).The first step - is lymphocytes representation. The first task is to determine how lymphocytes will be presented. Since the solution is the function (superposition of simple defined functions), then lymphocyte should represent this function encoded in some way. A tree of expression can be considered a convenient representation method (for implementation and for the further algorithm operation). As the number of variables is known in advance, lymphocyte also keeps a list of valid variables. Lymphocytes should support the following operations:
• Calculation of the value of the provided function at a point.
• Calculation of the affinity measure (discussed below) of the given lymphocyte.
• Simplification of the corresponding expression.
• Return of the string representation of an expression (with properly spaced brackets).
2) The second step - is affinity (the objective function). Lets introduce the concept of affinity. In biology the affinity of antibodies – is the binding strength of active sites of antibody molecule with determinant antigen groups. The antigen is a set of input data (function values of at the given points). In this model affinity characterizes the degree of approximation "success", i.e. is the value of the objective function
3) The third step- is the definition of network structure. The network is homogeneous and consists only of B-lymphocytes described above.

Lets consider the use of described symbolic regression algorithm for solving classification problem, which emerge by ecological forecasting [1]; the foundation of which - is the diffusion equation in the moving media: let the concentration of a pollutant in a certain region in a certain time is known. Lets assume that instead of the pollution sources capacity referred to in the problem formulation, we know the experimental observations data derived from a set of detectors. The problem of identification is to restore the function of the source, along this parallel is the approximate solution of the problem - solving function.

2 Problem formulation

Let us state the brief formulation of the symbolic regression problem [5]. There are many values of free variables $\{x_1,\ldots,x_r\},$, where $x_i \in R$, and corresponding values of the function $\{y_1,\ldots,y_r\}$.. These two sets form a set of baseline data *D*. Also, a set of functions is given that will be used in the construction of superposition. We will consider only continuously differentiable functions: $g: R^n \to R$, such as sin, cos, polynomials.

Let us consider the arbitrary superposition f, consisting of no more than m functions g. It is required to find such a superposition, which would provide the maximum (minimum) of the given functional $p(f, D)$.

This functional defines the objective function. It is chosen in such a way that it shows the degree of approximation of the built functions to the desired one.

Model of an artificial immune system consists of the lymphocytes representation and the algorithm of artificial immune system functioning. Lymphocytes are presented in the form of the expression trees. The solution - is the function (superposition of simple functions), therefore a lymphocyte should represent itself as a function encoded. A tree of expression can be considered a convenient representation for the realization, as well as for the subsequent operation of the algorithm. Since only binary and unary operations will be used, this tree will be binary. Function represented by a lymphocyte, can be written as

$$F = f_1(f_2(\ldots f_m(x_1, x_2, \ldots x_n))),$$

where $f_1,\ldots f_m$ are functions from the given set, *m* is the number of functions, less than or equal to the maximum permissible height of the expression tree, and x_i are free variables.

As the number of variables is known in advance, a lymphocyte also stores the list of feasible variables.

Lymphocytes should support the following operations:

• Calculation of value of the provided function at a point.

• Calculation of affinity measure of the given lymphocyte.

• Simplification of the corresponding expression.

• Return of a string representation of expression.

As the affinity (the objective function), let us use the following function: $\sqrt{\sum_{i=1}^{r}(f(x_i)-y_i)^2}$, , where $\{x_1,\ldots,x_r\}$, $x_i \in R^n$ is a given set of values of free variables; $\{y_1,\ldots,y_r\}$ is a given set of values of the unknown function at the corresponding points x_i , and $f(x_i)$, $i=\overline{1,r}$ is the range of function values represented by the given lymphocyte at the corresponding points x_i.The form of the function *U* after 200 iterations of an artificial immune system is shown in Fig. 2

3 The structure of artificial immune system

We suggest that the mathematical statement of artificial immune system could be presented as a set of the following elements [16]:

$AIS = \langle L, G, A, m, S \rangle$

where:

- *AIS* is artificial immune system;
- L is the space of all possible lymphocytes, the lymphocyte can represent a line, the list of coordinates, an expression tree;
- G is a set of all possible anti-genes, in so doing G can be the line, a matrix of logical values, and/or a list the list of values of a function in the known points; - $A: L \times G \to [0,1]$ is the given measure of affinity which assigns to each lymphocyte and each anti-gene a certain number from the segment [0, 1] (this number shows how "well" this lymphocyte reacts on the given anti-gen; $\mu: L \to L$ is the mutation operator , which is applied to a separate lymphocyte for the improvement of its recognition property; - $S: A \subset L \to B \subset A \subset L$ is the selection operator leaving the best lymphocytes in the current immune system, supporting the network size.

Then algorithm could be presented as a sequence of the following steps:

- Step 1 is to create initial immune system – $\mathrm{Im}\, System \subset L$.

On this step, the given number of admissible lymphocytes for a particular problem is generated randomly to form an initial system.

- Step 2 is to get

$g \in G, \forall l \in \mathrm{Im}\, System: a_l = A(l, g)$.

On this step, the corresponding antigens approach to all lymphocytes of the current immune system, and affinity (fitness) is calculated.

- Step 3 is to define the best lymphocyte $l^* = \arg\max(a_l)$.

- Step 4 is to apply the mutation operator $M = \{\mu(l), l \in \mathrm{Im}\, System\}$ to lymphocytes.

The mutation operator may be applied not to all lymphocytes, but to a certain subset (more frequently to those possessing the higher value of affinity). The mutation operator inserts little changes in the value or structure of a lymphocyte.

-Step 5 is to utilize the selection operator $\mathrm{Im}\, System = S(\mathrm{Im}\, System \cup M)$, which selects and remains lymphocytes with the greatest values of affinity from the current set of lymphocytes and from set of the mutated lymphocytes obtained on step 4.

- Step 6 is to decide whether *l** satisfies to the given criterion or whether the maximum number iteration is achieved. If yes, then go to the exit, otherwise return to Step 2.

Summarizing, it could be noted that the immune system solves a problem of function optimization that represents affinity. For various technical tasks this model and algorithm could be changed depending on the problem under consideration.

1. The first task is to define how the lymphocytes will be presented. Since we have a function as a solution of the problem, i.e. superposition of the defined simple functions, then the lymphocyte should be this function, encoded in a certain way. A convenient presentation of the lymphocytes for direct implementation, as well as for the further work of the algorithm, could be the expression tree. This tree will be binary by means of binary and unary operations. For example for the expression $x - 2 * (1/x + x/3)$ the tree will look like the following (Fig. 1) [4-7]:

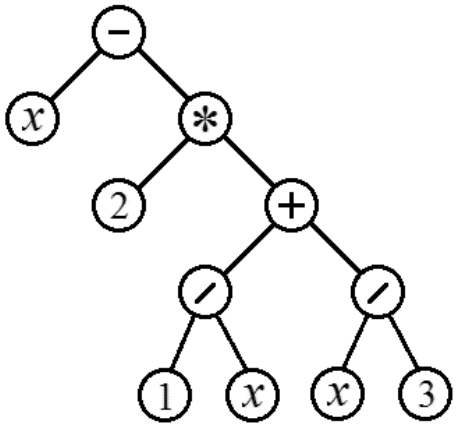

Fig. 1. The expression tree sample.

Since the number of variables is known beforehand, a lymphocyte also stores a list of valid variables.

All lymphocytes need to support the following operations:

- Calculation of the value of the given function at a point.
- Calculation of the affinity of the given lymphocyte (discussed below)
- Simplification of the corresponding expression.
- Return of the line representation of expression (with the correctly placed brackets)

Thus the function presented by the lymphocyte can be written as [10-11,14-16]:

$$F = f_1(f_2(...f_m(x_1, x_2,...,x_n))),$$

Where $f_1, f_2,...f_m$ – functions from the given set, m– number of functions, less than or equal to the maximum allowable height of an expression tree, xi – free variables.

2. Affinity (objective function). Let us introduce the concept of affinity. In biology, affinity of antibodies is the binding strength of the active centers of the antibody molecule with antigen determinant groups. The antigen is a set of input data - values of functions at given points. In the given model affinity characterizes the degree of "success" of approximation, i.e. represents the value of the objective function. Let us use the following function:

$\sqrt{\sum_{i=1}^{r}(f(x_i - y_i)^2}$,where: $\{x_1,...,x_r\}$, $x_i \in R^n$ – a specified set of values of free variables; $\{y_1,...,y_r\}$ – a specified set of values of the search function at appropriate locations x_i;

$f(x_i), i = \overline{1,r}$ – multiple values of the function represented by the given lymphocyte in corresponding locations x_i.

- The value of the function at a point would be calculated;
- The measure of affinity would be calculated;
- The corresponding expression would be simplified;
- The representation of expression would be returned with correctly placed brackets.

In immunology, affinity is a measure of how well selected B cell interacts with the antigen. In our model affinity is a measure of how close the function (represented by the lymphocyte) is to solution. We present the affinity function using the Euclidian metrics.

3. Immune system or immune network consists of many lymphocytes. Because of decentralized nature of the immune system, artificial immune systems can have many groups of lymphocytes, which can be placed on different computational nodes. These groups of lymphocytes can communicate with the others, share good solutions and maintain a variety of solutions.

4 Algorithms of AIS Realization

Let us describe a distributed immune system (fig. 3). The model p2p-interection [10] was chosen as a form of architecture of interaction of computing units. Peer-to-peer decentralized or peering (P2P - peer to peer) network is a computer network based on the equality of its participants. In this network there are no dedicated servers, and each node (peer) is both a client and a server. Unlike client-server architecture, this organization allows to maintain network performance with any number and any combination of available nodes.

In a distributed artificial immune system, each compute node will store some lymphocytes and at certain times share best lymphocytes with the neighboring nodes. Let us consider the application of the described algorithm of symbolic regression for the solution of classification problems arising in environmental forecasting. The model is taken from [1]. Its basis is the diffusion equation in the flowing system: let the concentration of the pollutant $\Phi = \Phi(x, y, t)$ in the area of $\Omega \subset R^2$ at the moment of time $t \in (0, T)$ satisfies the following initial-boundary value problem:

$$\frac{\partial \Phi}{\partial t} + u\frac{\partial \Phi}{\partial x} + v\frac{\partial \Phi}{\partial y} - \sigma\left(\frac{\partial^2 \Phi}{\partial x^2} + \frac{\partial^2 \Phi}{\partial y^2}\right) + \tau\Phi = P,$$
$$\Phi(x, y, 0) = \Phi_0;\ \Phi(x, y, t) = \Phi_1, \quad (x, y) \in \partial\Omega, \qquad (1)$$

where function $P = P(x, y, t)$

$$P(x, y, t) = \sum_{s=1}^{k} p_s(t)\delta_s(x, y; x_s, y_s)$$

characterizes the capacity of pollution sources.

5 Computational experiment

Suppose that $p_s(t)$ is the capacity of i sources of pollution distributed in the vicinity of (x_s, y_s) with the density $\delta_s(x, y; x_s, y_s)$; u, v are the components of the velocity vector; σ is a turbulence diffusivity coefficient; $\tau < 0$ is the parameter determining the absorption intensity of pollutant due to its entrainment, deposition, and other chemical reactions.

Let us perform a replacement

$$\Phi(x, y, t) = \exp(Ax + By + Ct)U(x, y, t),$$
$$P(x, y, t) = \exp(Ax + By + Ct)Q(x, y, t),$$

where $A = u/2\sigma$, $B = v/2\sigma$, $C = -\tau - (u^2 + v^2)/4\sigma$, then the differential equation can be converted into the heat equation

$$\frac{\partial U}{\partial t} - \sigma\left(\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2}\right) = Q. \qquad (2)$$

with initial-boundary conditions on the function U

$$U(x, y, 0) = U_0;\ U(x, y, t) = U_1, \quad (x, y) \in \partial\Omega.$$

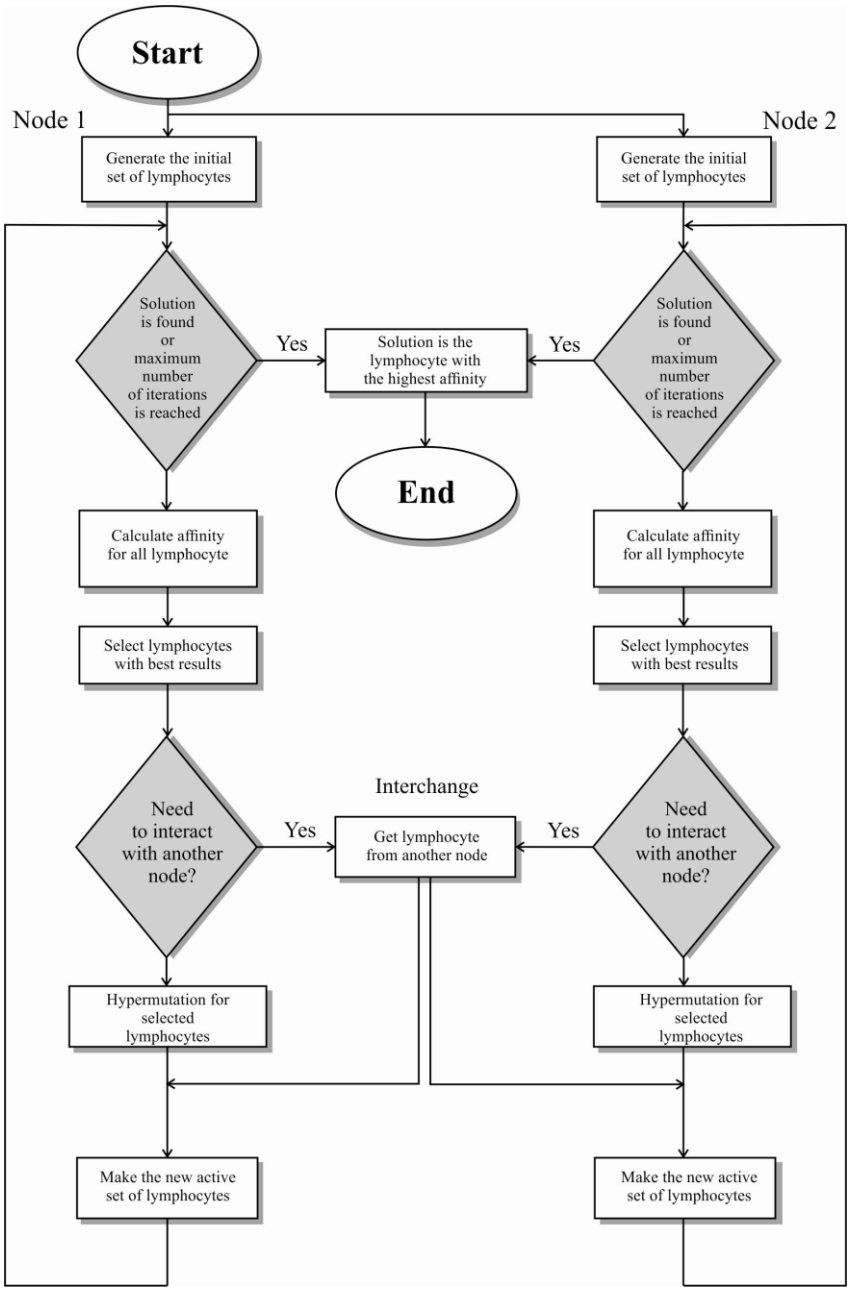

Fig.3. The algorithm of the distributed artificial immune system.

Let us consider this transformed problem assuming that instead of pollution sources the capacity $Q(x, y, t)$, referred in the direct problem formulation, the results of experimental observations are known $U(x_j, y_j, t_j) = \varphi_j$, $j = 1, \ldots, M$, derived from a set of sensors. The problem of identification is to restore function Q, along that exists the approximate solution of the problem in the form of function U.

Let us define the input data and the affinity function of the applied AIS $\{(x_j, y_j, t_j)\}_{j=1}^{M}$ – the sampling points from the area Ω, in which the values of the function are known $U(x_j, y_j, t_j) = \varphi_j$, $j = 1, \ldots, M$, The role of affinity will be performed by the following functional, showing how well the found function satisfies the initial-boundary value problem and the known observation results [12-15]:

$$J = \sum_{j=1}^{M} \left| \frac{\partial U}{\partial t} - \sigma \left(\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial x^2} \right) - Q \right|^2 (x_j, y_j, t_j) +$$

$$+ \lambda_0 \sum_{j=1}^{M} |U - U_0|^2 (x_j, y_j, 0) + \lambda_1 \sum_{j=1}^{M_1} |U - U_1|^2 (x_j, y_j, t_j) + \quad (3)$$

$$\lambda_2 \sum_{j=1}^{M} |U(x_j, y_j, t_j) - \varphi_j|^2.$$

This function is given piecewise: in the vicinity of the source, left and right, the lymphocyte will contain information about the respective three functions.

As a function that needs to be restored, we use the above function with the following parameters: $C = 1$, $c = 2$, $x_0 = 3$, $\delta = 1$, $\sigma = 2$.

Here $\{(x_j, y_j, t_j)\}_{j=1}^{M_1}$ are trial points in Γ; λ_0, λ_1, λ_2 are penalty parameters.

Since this problem is an approximate view of the two functions, the lymphocyte will consist of two expression trees that represent functions Q, U.

As an example the stationary solution in a one-dimensional case was used

$$U = \begin{cases} cx, & x \in (0; x_0 - \delta], \\ \dfrac{C-c}{4\delta}(x - x_0)^2 + \dfrac{C+c}{2} x - \dfrac{C-c}{2}\left(x_0 - \dfrac{\delta}{2}\right), & \\ \qquad\qquad x \in (x_0 - \delta; x_0 + \delta], & \\ Cx - (C-c)x_0, & x \in (x_0 + \delta; \omega). \end{cases} \quad (4)$$

As a function that needs to be restored, we use the above function with parameters $C = 1$, $c = 2$, $x_0 = 3$, $\delta = 1$, $\sigma = 2$. Then the first part is equal to

$$Q = \begin{cases} 0, & x \in (0; x_0 - \delta], \\ -\sigma \dfrac{C-c}{2\delta}, & x \in (x_0 - \delta; x_0 + \delta], \\ 0, & x \in (x_0 + \delta; \omega). \end{cases}$$

Since the function is given piecewise: in the vicinity of the source, left and right, the lymphocyte will contain information about the respective three functions.

Then

$$U = \begin{cases} 2x, & x \in (0; 2], \\ -\dfrac{1}{4}(x-3)^2 + \dfrac{3}{2}x + \dfrac{5}{4}, & x \in (2; 4], \\ x + 3, & x \in (4; 10). \end{cases}$$

$$Q = \begin{cases} 0, & x \in (0; 2], \\ 1, & x \in (2; 4], \\ 0, & x \in (4; 10). \end{cases}$$

The form of the function U after 200 iterations of artificial immune system is shown in Fig. 2

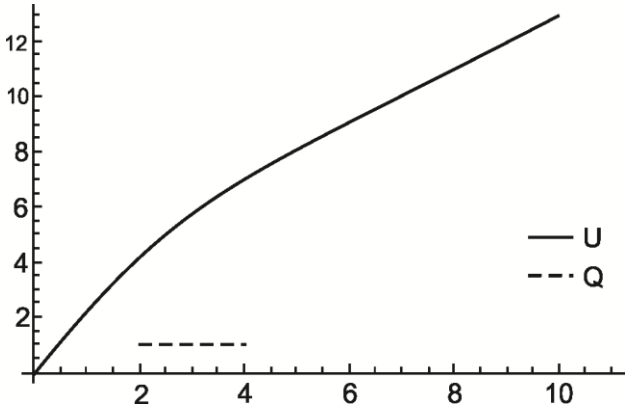

Fig. 2 Graph of the recovered functions U and Q

6 Realization

The main program consists of the following parts [16]:

1) Subsystem solution of the symbolic regression problem.

1.1 The module of a simple artificial system contains the classes of lymphocytes along with the methods to simplify expressions, to compute affinity and the algorithm of sequential implementation of the artificial immune system.

1.2 The module for the organization of the interaction between artificial immune systems in the network.

1.3 The module of implementation of a distributed artificial immune system.

1.4 The module of testing of the different functions of the subsystem, including the methods of computational experiment.

For the realization the Python language (version 3.3) has been selected [16]. This language has been chosen for the following reasons [16]:

• Cross-platform reason: during the creation of an application, running in a heterogeneous computing environment, the cross-platform factor is very important. Python allows you to create a program that does not require modification of the source code and recompiling even to work on a variety of platforms.

• The speed of development reason: Python is oriented on the increase of developer productivity and code readability. It also supports a variety of programming paradigms: structured, object-oriented, and functional [16].

• The standard library reason: includes a large number of useful functions and classes for various tasks such as networking, multithreading, process management.

During the creation process a system of version control Git has been utilized, with the source code stored on Github.

7 Conclusion

1.A unified model for the solution of identification problems is developed, where symbolic regression based on artificial immune system is using the fact that one of the main tasks of the immune system is to recognize harmful organisms and molecules.
2. A single algorithm implementing artificial immune system to solve identification problems in the symbolic regression is developed, along with its improved distribution version, using the property of immune system decentralization.
3. A software package, which includes the above-described algorithm and computational experiments, conducted for this proposed task, are developed.

Further on it might be possible to use immune systems to build programs for searching antivirus and organization of the software protection against unauthorized access. Artificial immune system might happened to be just a simple device in the organization of information security process.

References

[1] A.N. Vasilev, D.A. Tarhov Parametric neural network models of building of regularization of identify problems solution in ecology. *Contemporary information technologies and IT-education* **1** (2014).

[2] D. Dasgupta. *Artificial Immune Systems and their Application: collection of papers*. Moscow.: FIZMATLIT (2006).

[3]. D. Dasgupta, S. Yua, F. Nino. Recent Advances in Artificial Immune Systems: Models and Applications. *Appl. Soft Computing* **11** (2011).

[4] F. Freschi, M. Repetto. Multiobjective optimisation by a modified artificial immune system, *Artificial Immune Systems* (2005), 248–261.

[5] N.X. Hoai, R.I. McKay, D. Essam. Solving the symbolic regression problem with tree-adjunct grammar guided genetic programming: the comparative results. Evolutionary Computation, CEC '02. *Proceedings of the 2002 Congress* **2** (200) 1326-1331.

[6] I. E. Hunt, D. E. Cooke. Learning using an artificial immune system. *J. Network Comp. Appl.* **19** (1996) 189-212.

[7] A.A. Ishiguro, Y. Watanabe, T. Kondo. Robot with a decentralized consensus-making mechanism based on the immune system. *Proceedings of ISADS* (1997) 231-237.

[8] C.G. Johnson. Artificial immune systems programming for symbolic regression. *Genetic Programming: 6th European Conference* (2003) 345–353.

[9] J.O. Kephart. A biologically inspired immune system for computers. *Proceedings of Artificial Life IV: The Fourth International Workshop on the Synthesis and Simulation of Living Systems* (1994) 130–139.

[10] R.A. Schollmeier. Definition of peer-to-peer networking for the classification of peer-to-peer architectures and applications. *Proceedings of the First International Conference on Peer-to-Peer Computing, IEEE* (2001) 101-102.

[11] E. Hart, J. Timmis. Application areas of AIS: The past, the present and the future. *Appl. Soft Computing* **8** (2008) 191–201.

[12] M. Bardeen. Survey of methods to prevent premature convergence in evolutionary algorithms. Workshop of Natural Computing, *J. Chilenas de Computacion* (2013) 13–15.

[13] D. Barkai. Peer-to-Peer Computing. Santa Clara: Intel Press (2002).

[14] K. Bennett, M.C. Ferris, Y.E. Ioannidis. A genetic algorithm for database query optimization. *Proceedings of the fourth International Conference on Genetic Algorithms* (1991) 400-407.

[15] H. Bersini H. The endogenous double plasticity of the immune network and the inspiration to be drawn for engineering artifacts.

Artificial Immune Systems and Their Applications (1999) 22-44.
[16] I.F. Astakhova, S.A. Ushakov, Ju.V. Hitskova. Model and algorithm of an artificial immune systems for the recognition of single symbols and their comparison with existing methods. *WSEAS Trans. on Information Science and Applications* **13** (2016) 38-45.
[17]J.R.Koza. *Genetic Programming*. Cambridge: MIT Press (1998). — 220 p.
[18] W.B. Langdon , R. Poli.. *Foundations of Genetic Programming* Haidelberg: Springer-Verlag (2002).