Original article
Dependability, vol. 19 no.2, 2019 *DOI: 10.21683/1729-2646-2019-19-2-22-27*

Neural network integration of classical statistical tests for processing small samples of biometrics data

Alexander I. Ivanov, *Penza Research and Design Electrical Engineering Institute, Russian Federation, Penza*
Evgeny N. Kuprianov, *Penza State University, Russian Federation, Penza*
Sergey V. Tureev, *Research and Design Institute for Communications and Control Systems***,** *Russian Federation, Moscow*


Alexander I. Ivanov


Evgeny N. Kuprianov


Sergey V. Tureev

Abstract. *The* **Aim** *of this paper is to increase the power of statistical tests through their joint application to reduce the requirement for the size of the test sample.* **Methods.** *It is proposed to combine classical statistical tests, i.e. chi square, Cram r-von Mises and Shapiro-Wilk by means of using equivalent artificial neurons. Each neuron compares the input statistics with a precomputed threshold and has two output states. That allows obtaining three bits of binary output code of a network of three artificial neurons.* **Results.** *It is shown that each of such criteria on small samples of biometric data produces high values of errors of the first and second kind in the process of normality hypothesis testing. Neural network integration of three tests under consideration enables a significant reduction of the probabilities of errors of the first and second kind. The paper sets forth the results of neural network integration of pairs, as well as triples of statistical tests under consideration.* **Conclusions.** *Expected probabilities of errors of the first and second kind are predicted for neural network integrations of 10 and 30 classical statistical tests for small samples that contain 21 tests. An important element of the prediction process is the symmetrization of the problem, when the probabilities of errors of the first and second kind are made identical and averaged out. Coefficient modules of pair correlation of output states are averaged out as well by means of artificial neuron adders. Only in this case the connection between the number of integrated tests and the expected probabilities of errors of the first and second kind becomes linear in logarithmic coordinates.*

Keywords: *statistical tests: chi square, Cram r-von Mises, Shapiro-Wilk; artificial neural networks, small samples, normal law of data distribution hypothesis testing.*

For citation: *Ivanov AI, Kuprianov EN, Tureev SV. Neural network integration of classical statistical tests for processing small samples of biometrics data. Dependability 2019;2: 22-27. DOI: 10.21683/1729-2646-2019-19-2-22-27*

The problem of control of the data distribution law of small samples

The problems of ensuring the reliability of unique critical systems [1, 2] are multifaceted and can be solved only through a set of organizational and technical measures. These problems are especially prominent in neural network biometrics. Each of us has a unique biometric image that is to be transformed into a cryptographic key or long access password generated through random symbols. The uniqueness of the transformation is enabled by means of neural network learning, while the learning sample has a close to normal multidimensional data distribution law. The problem is that learning samples are small. In particular, the standard learning algorithm [3] is able to solve the task on samples of 20 examples, if this sample is obtained correctly and has no outliers (gross errors).

In cases of large biometrics data samples (200 tests and more) it is not difficult to test the hypothesis of normal distribution. The chi square criterion or any other statistical criterion can be used [4]. One of the problems of biometrics [5] is that its users do not wish to provide to an automatic neural network learning machine [3] 200 and more instances of their biometric image. Users feel satisfied having submitted a learning sample consisting of 10 to 20 examples of their unique biometric image, for example, a handwritten password or voice password. Users perceive negatively the requirements to present more than 20 examples.

The situation is similar in botany, biology, and medicine. A plan breeder or a biologist is not able to quickly get a sample of 200 animals (plant specimens) with necessary rare characteristics. A sufficient sample for correct statistical estimation can be obtained after a long period of time by selecting and consolidating the desired rare characteristics over several generations.

There is a similar situation in medicine. Large samples are required to test statistical hypotheses. The subject matter of statistical processing of small samples is very popular, but the well-known recommendations [6, 7] do not significantly improve the situation. As a rule, improvements are achieved through the application of several statistical criteria [8].

An attempt could be made to enhance the known statistical criteria [9], but this does not result in major improvements. As a rule, new statistical criteria or variants of earlier criteria individually provide poor results.

The main idea of this paper is the neural network integration of standard statistical criteria [4, 10, 11]. The progress achieved by the Russian neural network biometrics is very significant. Regulators of the Russian information security market have developed the GOST R 52633.xx Russian national series of standards that regulate a number of tough requirements for neural network biometrics. In this paper we will actually attempt to apply the well-developed mathematical techniques of neural network biometrics to new subject areas. At the same time, we will try to show that the very tough requirements of the Russian information security regulators for the probability of error of the first and second kind can be fulfilled in other subject areas through the implementation of the primary recommendations of the GOST R 52633.xx series of neural network biometrics standards.

Synthesis and adjustment of the chi square neuron with 5 inputs

When testing the normality hypothesis in practice, the Pearson's chi square test is most often used. For a small sample with 21 tests, the formula for calculating the chi square criterion value is the following:

$$\chi^2 = 21 \cdot \sum_{i=1}^{5} \frac{\left(\frac{n_i}{21} - \Delta\tilde{P}_i\right)^2}{\Delta\tilde{P}_i}, \qquad (1),$$

where n_i is the number of tests in the i-th histogram interval; $\Delta\tilde{P}_i$ is the expected probability of tests being within the i-th histogram interval under the normal data distribution law of the checked sample.

Let us note that in accordance with the national standard recommendations [10], the average number of tests within each of the histogram intervals is to be close to 5. That is the reason why in formula (1) summation over 5 histogram intervals for a small sample of 21 tests is used.

When developing the formula in 1990, Pearson had no access to computer technologies. For this reason, he was forced to look for asymptotic relations for infinitely large samples. Today the situation has changed. Any student is able to write a program that can produce millions of samples of 21 tests. Figure 1 shows the probability density distribution of the chi square criterion values for samples with a normal and uniform value distribution law.

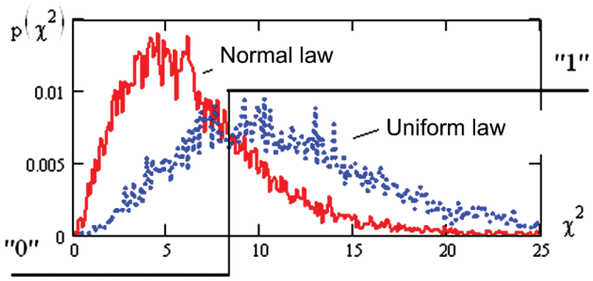

Figure 1. Distribution of chi square criterion values for samples with a normal and uniform value distribution law

Let us note that artificial neurons are configured in such a way as to effectively divide input data into two classes: normal and uniform [12]. Figure 1 shows that the threshold element of the chi square neuron divides the continuum of output elements into two areas: 0 is normal data and 1 is uniform data. The output quantifier of a chi square neuron is configured based on the condition of equally probable error values of the first and second kind of $P_1 = P_2 = P_{EE} = 0.292$.

Let us sort the data of the checked sample according to their values to obtain five input parameters of chi square neuron:

$$x = \text{sort}(x). \tag{2}$$

At the same time, it is required to calculate the width of the histogram intervals:

$$\Delta x = \frac{x_{20} - x_0}{5}. \tag{3}$$

Furthermore, the position of the interval ends is calculated:

$$X_i = x_0 + \Delta x \cdot i \quad \text{when} \quad i = 0, 1, ..., 5. \tag{4}$$

Only after that, it is possible to calculate the number of hits for each of the histogram intervals and form a vector of input parameters $\{n_1, n_2, \ldots, n_5\}$ for the neuron (1). The final result is quantized:

$$\begin{cases} z(\chi^2) \leftarrow \text{"0"} \ \text{ if } \ \chi^2 \le 7.72; \\ z(\chi^2) \leftarrow \text{"1"} \ \text{ if } \ \chi^2 > 7.72. \end{cases} \tag{5}$$

As the result, we have a complete formal description of the chi square neuron implementation for a sample of 21 tests.

Synthesis and configuration of Shapiro-Wilk neuron with 10 inputs

Obviously, the Shapiro-Wilk criterion can be applied to the same sample of 21 tests [4, 11]. This criterion is calculated as following:

$$\upsilon^2 = \frac{1}{\left(\sigma(x)\right)^2} \cdot \left\{ \sum_{i=0}^{9} a_i \cdot \left(x_{20-i} - x_i\right) \right\}^2, \tag{6}$$

where x_i is the ordered values of the sample being checked, $\sigma(x)$ is the standard deviation, a_i is the table values of the Shapiro-Wilk coefficients.

Figure 2 shows the distribution of the values of this criterion for the uniform and normal laws.

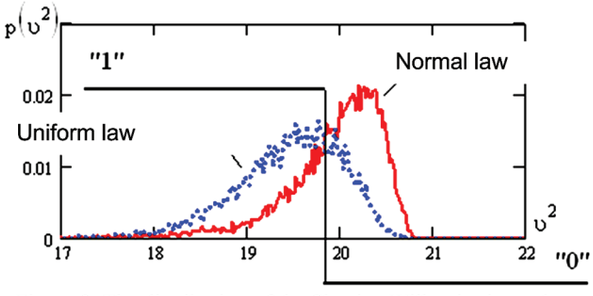

Figure 2. The distribution of the Shapiro-Wilk criterion values for the samples with 21 tests with uniform and normal distribution laws

If the functions of (6) are considered as some kind of artificial neuron, then its outputs will be 10 differences of data of the sample being checked, and the output quantifier will be described as follows:

$$\begin{cases} z(\upsilon^2) \leftarrow \text{"0"} \ \text{ if } \ \upsilon^2 \ge 19.8; \\ z(\upsilon^2) \leftarrow \text{"1"} \ \text{ if } \ \upsilon^2 < 19.8. \end{cases} \tag{7}$$

Such configuration of the threshold of the quantifier provides the errors probability of the first and second kind of $P_1 = P_2 = P_{EE} = 0.303$.

Synthesis and configuration of a Cram r-von Mises neuron with 20 inputs

If we compare the chi square neuron (1) and the Shapiro-Wilk neuron (6), we can see the growth of their input dimension (the number of inputs of their adders). The Cramér-von Mises neuron has an even higher input dimension:

$$\omega^2 = \sum_{i=0}^{19} \left(\frac{i+1}{21} - \tilde{P}(x_i) \right)^2 \cdot \frac{x_{i+1} - x_i}{x_{20} - x_0}. \tag{8}$$

Figure 3 shows the distribution of values at the output of the Cramér-von Mises neuron adder.

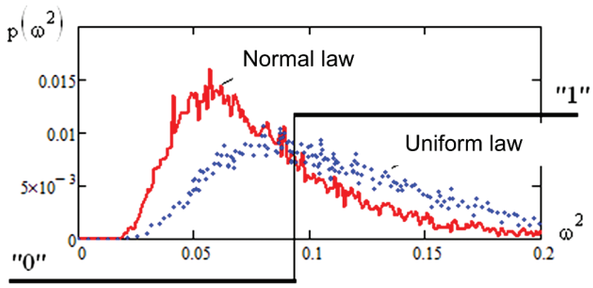

Figure 3. Distribution of values of the Cramér-von Mises criterion for samples with 21 tests with the uniform and normal distribution laws

The configured output quantifier of the neuron adder is described as follows:

$$\begin{cases} z(\omega^2) \leftarrow \text{"0"} \ \text{ if } \ \omega^2 \le 0.087; \\ z(\omega^2) \leftarrow \text{"1"} \ \text{ if } \ \omega^2 > 0.087. \end{cases} \tag{9}$$

Such threshold configuration for quantifier operation provides the same values of errors probability of the first and second kind of $P_1 = P_2 = P_{EE} = 0.342$.

Joint application of three statistical criteria

The statistical criteria described above are linearly independent (they have modules of correlation coefficients less than 1):

$$\begin{cases} \text{corr}(\chi^2, \upsilon^2) \approx +0,559; \\ \text{corr}(\chi^2, \omega^2) \approx -0,708; \\ \text{corr}(\omega^2, \upsilon^2) \approx -0,667. \end{cases} \tag{10}$$

The absence of a complete linear dependence (10) of the output states of the three criteria allows combining them for joint application. In this case, the output code of the three neurons "000" will correspond to a triple confirmation of the hypothesis of the data normality of the checked sample. The inverted state of this code "111" will correspond to the triple confirmation of the hypothesis of the uniform law of distribution of small sample data.

Let us consider one of two hypotheses for the majority of states of "0" or "1" in the output code of the three neurons code by analogy with practical application of neural network converters, which is biometrics and code. In this case, each of the four code states "normal distribution" will correspond to its own probability of errors. Table 1 shows these data.

Table 1. Error probability for the code states "normal distribution"

Code	"000"	"001"	"010"	"100"
P_1	0.0404	0.0423	0.0441	0.0621

Then, if we consider the codes from Table 1 as some complex characteristic of "data normality" it can lead to errors arising with the probability from 0.0404 to 0.0621. There is about a 7-fold decrease in probability of taking wrong decisions, when using three statistical criteria in comparison with their application one by one.

Effect of increasing accuracy of estimates with the growing size of the group of neural network integration of statistical criteria

Dozens of statistical criteria have so far been developed and applied [4, 10, 11]. Supposedly, an equivalent artificial neuron can be developed for each of them. Moreover, previously unknown statistical criteria are under development [13–17]. The first progress in this area will allow adding dozens of completely new statistical criteria to the existing ones. That means that in a few years it will be possible to develop a series of hundreds of different statistical criteria and their neural analogs.

The question arises: up to what level is it possible to reduce the probability of errors by means of neural network integration of a collection composed of 100 and more statistical criteria? This question can be answered based on the accumulated technological experience in processing of neural network biometrics data.

The neural network symmetrization technology can be used for prediction [18, 19]. To implement it, let us average the error probability of the three previously examined neurons (0.292+0.303+0.342)/3 = 0.312. Then, let us average the modules of correlation coefficients between the output states of the three neurons (10): $E(|\text{corr}(.)|) = 0.645$. We proceed from the fact that all of the 100 integrated criteria have symmetric matrices of correlation coefficients with the elements' values outside its diagonal of 0.645.

Another simplification is the normalization of the output states of neuron adders that contradicts the data presented in Figures 1, 2, 3, but at the moment only for this simplification there is a positive experience of using symmetrization.

Figure 4 shows the block diagram of the numerical experiment. Initial data for the numerical experiment are obtained from 100 software generators of pseudorandom numbers with normal distribution.

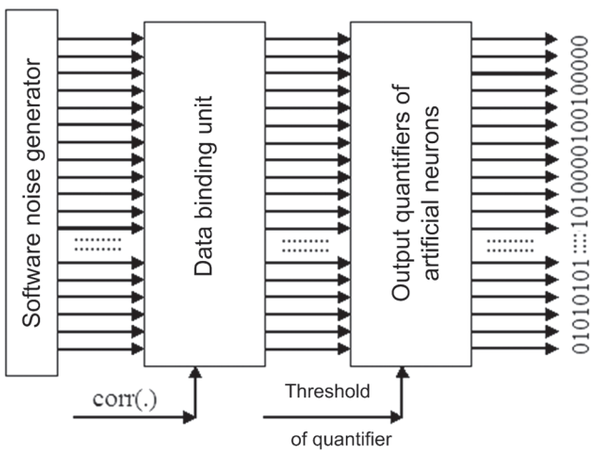

Figure 4. Block diagram for modelling completely symmetrical artificial neural networks

As 100 software generators provide independent data, such data needs to be interconnected and correlated equally. Figure 4 shows that this function is performed by the second left block that multiplies the vector of independent random numbers and by a symmetric connecting matrix:

$$\begin{bmatrix} 1 & a & \vdots & a \\ a & 1 & \vdots & a \\ \vdots & \vdots & \vdots & \vdots \\ a & a & \vdots & 1 \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{100} \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{100} \end{bmatrix}. \quad (11)$$

Due to the symmetry that connects the transfer matrix (11), the output data is correlated equally. To obtain a given value of coefficients of equal correlation $\text{corr}(y_i y_{i+1}) = 0.645$, it suffices to find the value of one control parameter a.

Let us note that the procedure of relations and data symmetrization cannot provide exact correspondences of predictions and real data. If we set the quantization threshold of the neuron emulation block in such a way that the error probability is 0.312, then the output triple will have a total error of 0.138. This result is about 3 times worse than the actual data in Table 1.

Let us reduce the equal probability of errors of each neuron from 0.312 to 0.141 to match the results with the observed data. In this case the probability of errors of joint operation of three neurons will be 0.0404.

The transition from normal data to data with the equal correlation is profitable as for this special case in logarithmic coordinates the error probabilities and number of neurons are connected by a linear dependence as shown in Figure 5.

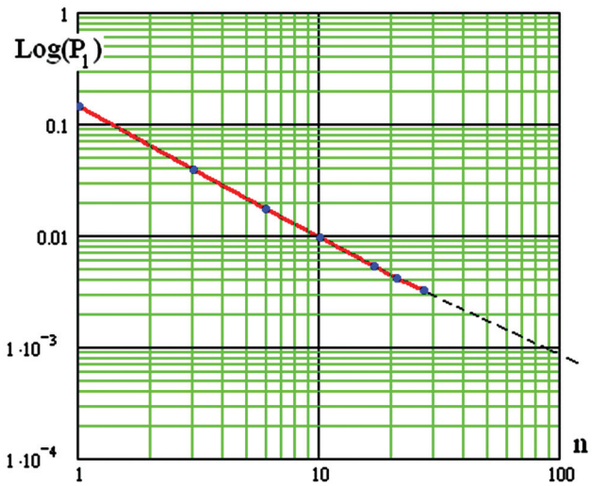

Figure 5. The line of decreasing probability of errors of the first kind due to application of several statistical criteria with correlation coefficients of 0.645

The line was constructed in 7 groups, composed of 1, 3, 6, 10, 16, 21, 27 neurons. When conducting an experiment, a sample of 10 000 000 tests was used; the computation time for a conventional computer is about 9 minutes. It should be noted that using this computer it is difficult to conduct a numerical experiment for a group of 100 neurons, as such experiment would take several months. It is possible to reduce the time by means of extrapolation (dashed line in Figure 5).

As the result, the predicted value of the error probability for a neural network generalization of 10 criteria should be $P_1 = 0.01$, and when summarizing 100 criteria the probability of errors should go down to 0.0009. Such a significant reduction of errors probability is a greater incentive for the synthesis of new statistical criteria [13-17].

Conclusion

Pearson, who created the chi square criterion in 1900, essentially launched a revolution in statistical processing. The path of development discovered by Pearson proved to be very fruitful and over 119 years his followers have created dozens of different statistical criteria.

Neural networks have been a focus of scientific research since the middle of the 20-th century, but only at the beginning of the 21-st century this technology was implemented into the industry and standardized [3].

The key statement of this paper is that it is possible to combine two seemingly different branches of mathematics. Their integration only requires the neural network biometric data processing technologies that are standardized in Russia and are applied to 3 or more standard statistical criteria. In the case of the considered triple of statistical criteria, this approach reduces the probability of errors more than 7 times. In this case, thesis on expediency of expansion of nomenclature of the existing statistical criteria becomes obvious. The larger is the size of the group of statistical criteria generalized by neurons the better is the final result.

In this context, the approach to the synthesis of new statistical criteria is fundamentally changing. After Pearson, mathematicians were trying to find a new, more powerful criterion. A great number of analyzed criteria proved to have low power, and therefore were not published. With neural network integration of a set of statistical criteria, the power of each of them becomes secondary. The correlation relationships between the added criterion and the group of other criteria are also very important. In our case, two integrated criteria have almost the same power, but in this group there is a special Shapiro-Wilk criterion that has low correlation with the primary chi square and Cramér-von Mises criteria.

Thus, the possible diversity of statistical criteria is to be researched again, taking into account not only their relative power, but also the values of their correlation coefficients in groups with other relevant statistical criteria. New statistical criteria with relatively low power of hypothesis separation were previously rejected and not published, but now the situation has changed. It is more important to understand how the new criterion complements the already studied statistical criteria. Probably, it will be necessary to create a table of the level of affinity (correlation) of already known and promising statistical criteria in the nearest future. Linearly independent (weakly correlated) statistical criteria have to be grouped, and neural network integrations are to be created for them.

References

[1] Pokhabov Yu.P. Problems of dependability and possible solutions in the context of unique highly vital systems design. Dependability 2019;19(1):10-17.

[2] Pokhabov Yu.P. Ensuring dependability of unique highly vital systems. Dependability 2017;17(3):17-23,

[3] GOST R 52633.5-2011. Information protection. Information protection technology. The neural net biometry-code converter automatic training. Moscow: Standartinform; 2012 [in Russian].

[4] Kobzar A.I. Prikladnaia matematicheskaia statistika: dlia inzhenerov i nauchnykh rabotnikov [Applied mathematical statistics: for engineers and researchers]. Moscow: FIZMATLIT; 2006 [in Russian].

[5] Iazov Yu.K., editor, Volchikhin V.I., Ivanov A.I., Funtikov V.A., Nazarov I.G. Neyrosetevaia zashchita personalnykh biometricheskikh dannykh [Neural network protection of biometric data]. Moscow: Radiotekhnika; 2012 [in Russian].

[6] Sukhoruchenkov B.I. Analiz maloy vyborki. Prikladnye statisticheskie metody [Small sample analysis. Applied statistical methods]. Moscow: Vuzovskaya kniga; 2010 [in Russian].

[7] Doerffel K. Statistics in analytical chemistry. Moscow: Mir; 1994.

[8] Dayev Zh.A., Nurushev E.T. Application of statistical criteria for improving the efficiency of risk assessment methods. Dependability 2018;2:42-45.

[9] Akhmetov B.B., Ivanov A.I. Estimation of quality of a small sampling biometric data using a more efficient form of the chi-square test. Dependability 2016;16(2):43-48.

[10] R 50.1.037-2002. Rekomendatsii po standartizatsii. Prikladnaia statistika. Pravila proverki soglasia opytnogo raspredelenia s teoreticheskim. Chast I. Kriterii tipa $\chi 2$ [Standardization recommendations. Applied statistics. Rules for compliance verification of experimental distribution with the theoretical one. Part I. Chi-square-type criteria]. Moscow: Gosstandart of Russia; 2001 [in Russian].

[11] R 50.1.037-2002. Prikladnaia statistika. Pravila proverki soglasia opytnogo raspredelenia s teoreticheskim. Chast II. Neparametricheskie kriterii [Applied statistics. Rules for compliance verification of experimental distribution with the theoretical one. Part II. Non-parametric tests]. Moscow: Gosstandart of Russia; 2002 [in Russian].

[12] Haykin S. Neural Networks: A Comprehensive Foundation. Moscow: Viliams; 2006.

[13] Serikova N.I., Ivanov A.I., Serikova Yu.I. Otsenka pravdopodobia gipotezy o normalnom raspredelenii po kriteriu Dzhini dlia chisla stepeney svobody, kratnogo chislu opytov [Likelihood estimation of the hypothesis of normal Gini distribution for the number of degrees of freedom multiple of the number of experiments]. Voprosy radioelektroniki 2015;1(1):85-94 [in Russian].

[14] Perfilov K.A. Kriteriy srednego geometricheskogo, ispolzuemyy dlia proverki dostovernosti statisticheskikh gipotez raspredelenia biometricheskikh dannykh [Criterion of geometric mean used for validity verification of the statistical hypotheses of biometric data distribution]. Proceedings of the science and technology conference of the Penza information technology security cluster. Penza; 2014. Vol. 9. p. 92-93, <http://www.pniei.penza.ru/RV-conf/T9/S92> [in Russian].

[15] Ivanov A.I., Perfilov K.A. Otsenka sootnoshenia moshchnostey semeystva statisticheskikh kriteriev "srednego geometricheskogo" na malykh vyborkakh biometricheskikh dannykh [Assessment of the relations of the strengths of the "geometric mean" family of statistical tests on small samples of biometric data]. XI All-Russian Science and Practice Conference Modern security technologies and comprehensive facility security assets. Penza-Zarechny; 2016. p. 223-229 [in Russian].

[16] Ivanov A.I., Perfilov K.A., Malygina E.A. Multivariate statistical analysis of the quality of biometric data on extremely small samples using the criteria of the geometric mean tests calculated for the analyzed probability functions. Measuring. Monitoring. Management. Control 2016;2(16):64-72 [in Russian].

[17] Ivanov A.I., Perfilov K.A., Malygina E.A. Evaluation of the quality of small samples of biometric data using a differential variant statistical test of the geometric mean. Vestnik SibGAU 2016;4(17):864-871 [in Russian].

[18] Malygin A.Yu., Volchikhin V.I., Ivanov A.I., Funtikov V.A. Bystrye algoritmy testirovania neyrosetevykh mekhanizmov biometriko-kriptograficheskoy zashchity informatsii [Fast algorithms for testing of neural network mechanisms of biometric and cryptographic protection of information]. Penza, Penza State University Publishing 2006 [in Russian].

[19] Akhmetov B.S., Volchikhin V.I., Ivanov A.I., Malygin A.Yu. Algoritmy testirovania biometriko-neyrosetevykh mekhanizmov zashchity informatsii [Algorithms for testing biometric and neural network mechanisms of information protection]. Almaty: Satpayev KazNTU; 2013 [in Russian].

About the authors

Alexander I. Ivanov, Doctor of Engineering, Associate Professor, Lead Researcher of Laboratory of Biometric and Neural Network Technologies, Penza Research and Design Electrical Engineering Institute, Russian Federation, Penza, e-mail: ivan@pniei.penza.ru

Evgeny N. Kuprianov, post-graduate student, Department of Information Security Technology, Penza State University, Russian Federation, Penza, e-mail: ibst@pnzgu.ru

Sergey V. Tureev, Head of Research and Technology Center, Research and Design Institute for Communications and Control Systems, Russian Federation, Moscow, e-mail: niissu@niissu.ru

Received on: 22.01.2019