Journal of Applied Engineering Science

doi:10.5937/jaes16-18435 Paper number: 16(2018)2, 546, 404 - 409

A SIMPLE WAY TO ASSESS THE SPECTRAL LINES INFORMATIVITY OF A CHI-SQUARE MOLECULE IN ANALYZING SMALL SAMPLES OF BIOMETRIC DATA

Alexander Ivanov[1*], Alexei Gazin[2], Yulia Serikova[3]

[1]Penza Scientific Research Electro-Technical Institute, Russian Federation
[2]Lipetsk State Pedagogical P. Semenov-Tyan-Shansky University, Russian Federation
[3]Penza State University, Russian Federation

The prerequisites for reducing the test sample chi-square Pearson test size from 400 to 32 or fewer examples while maintaining its power are considered. The urgency of the problem results from the fact that when learning and testing the biometric identification means to identify the personality, it is not possible to use large volumes of learning and test samples. The conditions under which the chi-square test on small samples from the continuous distribution of values becomes a discrete distribution of values are formalized. Normal and uniform laws of values distribution use histograms with uniform intervals, which accurately relate the central intervals of the histogram to the mathematical expectation calculated on the test sample. 16 experiments shown that the chi-square-synchronized test built on histograms with four equal intervals has a discrete probability spectrum consisting of only 20 significant spectral lines. A simple method for estimating the informativity of each of the important spectral components is proposed. Traditional statistical assessments can be strengthened by the following deeper level of the spectral components analysis of small samples of biometric data. The second deeper level of statistical processing should be substantially more powerful. Under the same conditions, the computational informativity increases from 2.22 bits to 24.95 bits due to the transition from simple continual calculations to discrete calculations of high computational complexity.

Key words: Chi-square Pearson test, Small samples of biometric data, Quantum effects of continuums representation by small samples, Distributivity, Histogram, Spectral lines

INTRODUCTION - THE PROBLEM OF PERFORMING STATISTICAL ASSESSMENTS ON SMALL SAMPLES OF BIOMETRIC DATA

In practice, data is always low in amount. This problem is especially acute when solving the problem of biometric identification of an individual [01, 02, 03]. A person is a very complex object of high dimensionality, which significantly complicates statistical multidimensional estimates. The same problems arise in the transition from a person's personal biometrics to collective biometrics of groups of people. For example, when testing new drugs, it is required to involve a sufficiently large number of biometric data of patients with the necessary pharmacists' disease. Then they should be treated, and then, having typed the necessary amount of statistics, confirm the harmlessness of the new pharmacological drug. This approach to testing new drugs usually takes several years.

Obviously, other things being equal, a two or three times reduction in the amount of the test sample can reduce the time of testing the biometric data two to three times and, accordingly, shorten the time for the withdrawal of pharmaceutical preparations to the market. That is, the correct reduction in the test samples volume is one of the effective ways to reduce the cost of drugs and reduce the time of their withdrawal to the market.

Statistical estimates often use the chi-square Pearson test because it knows the analytical description of the asymptotic distribution of values:

$$p(\chi^2(x,\mathrm{m})) = \frac{1}{2^{\frac{m}{2}} \cdot \Gamma\left\{\frac{m}{2}\right\}} \cdot x^{\left(\frac{m}{2}-1\right)} \cdot \exp\left(\frac{-x}{2}\right) \qquad (1)$$

where m is a positive integer of degrees of freedom, calculated through the number of columns of the histogram - k from the formula m = k-3, is the Euler gamma function.

Standardized recommendations for the application of the chi-square test [04] require that for each column of the histogram an average of about 5 or more samples be counted, and the sampling itself would be from 200 to 400 experiments. In biometrics, it is not always possible to obtain such a large sample. In connection with this, the actual task is to reduce the volume of the test sample to 16 examples while maintaining the power of the Pearson chi-square test. Naturally, the asymptotic relation (1) ceases to work with such small samples and other mathematical constructions must be created in its place.

QUANTUM EFFECTS ARISING FROM THE REPLACEMENT OF STATISTICAL CONTINUA BY SMALL SAMPLES OF BIOMETRIC DATA

Let us suppose that there is a samples set of the normal distribution of values, containing 16 experiments. Next let us build a histogram of data, consisting of 4 columns. The columns width of the histogram will be chosen acording to the following rule:

* *Penza Scientific Research Electro-Technical Institute, Russian Federation, alexander.ivanov.04@bk.ru*

$$\Delta x = \frac{\max(x) - \min(x)}{4} \tag{2}$$

Obviously, the histogram synthesizing operation is nothing else than an operation of quantizing a continuous normal continuum by a quantizer with 4 output states. Naturally, such a quantizer generates quantization noise of a continuous normal distribution. This situation is shown in Figure 1 below.

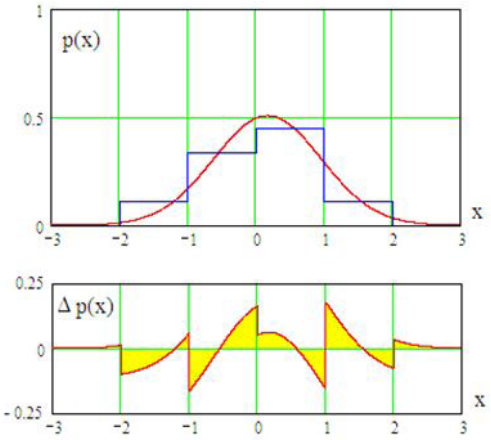

Figure 1: The effect of quantizing the continuous density of the values distribution by representing it with a histogram (upper part) and quantization noise (the lower part of the Figure)

Obviously, quantization noise causes the ratio (1) to stop working. It is also obvious that it is possible to improve the situation, for example, by smoothing the quantization noise [05, 06, 07] with a simple digital filter. In this case, the selection of the parameters of the digital smoothing filter will never allow for approaching the relation (1), because the smoothing filtering will lose the independence of the given chi-square functionals [08].

NUMERICAL MODELLING OF CHI-SQUARE DISTRIBUTION LAWS FOR SMALL SAMPLES

There were no powerful computers at the Pearson's times. Today the situation is different; it is possible to use the software generator of many samples from 16 experiments. Further it is possible to calculate a set of chi-square values from this data:

$$\chi^2 = 16 \cdot \sum_{i=1}^{4} \frac{\left(\frac{b_i}{16} - P_i\right)^2}{P_i} \tag{3}$$

where b_i is the number of experiments that hit the *i*-th interval of the histogram, P_i is the expected theoretical probability of falling into the *i*-th interval of the histogram under the normal distribution law.

Further, it is possible to construct the densities of chi-square values for the normal law of pseudo-random numbers and the uniform law of the original data. As a result, two curves are obtained for a million implementations, shown in Figure 2.

It can be seen from Figure 2 that the final density distributions of the values are non-monotonic. There are significant periodic bursts superimposed on some smooth chi-square distributions with different number of freedoms. It is obvious that the number of freedoms for normal initial data will be much less than the analogous number of freedoms for uniform initial data.

There is an illusion that the imposed oscillations are of a random nature and can be eliminated by increasing the volume of the number, accounted for the implementation of samples of 16 experiments. One can repeat the experiments and one will get the same result. It is possible to increase the number of realizations in tens or even hundreds of times, however, the vibrational components do not decrease their amplitude. All this testifies to the non-random (deterministic) appearance of the vibrational components of the distribution density values of Figure 2.

Vibrational superimpositions on smooth distribution functions of values are a consequence of the periodic noise quantization. The quantization noise envelope (the lower part of Figure 1) is periodic; it is the periodic nature of the quantization noise that leads to the appearance of a periodic component of the distribution density of Figure 2. Due to the partial dissynchronization of the periodic components of the quantization noise, there is a slight attenuation of the amplitude when data is accumulated, but complete elimination of the vibrational components for small samples of 16 examples never occurs.

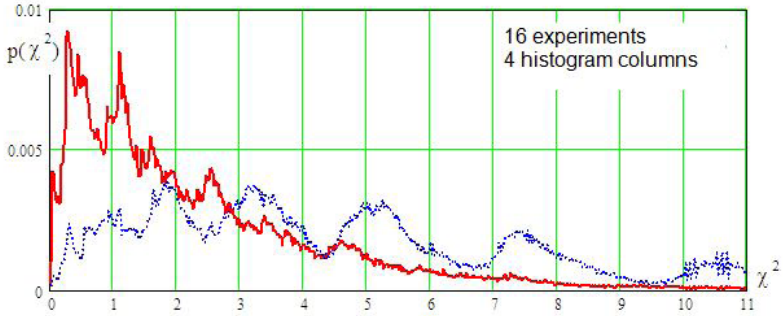

Figure 2: Chi-square distribution functions for normal source data (continuous line) and for raw data with a uniform law of distribution of values (dashed line)

At the weakening point of the periodic components of the quantization noise [05, 06, 07], averaging can be done differently and emphasize the periodic components present in the quantization noise [09, 10, 11, 12, 13].

Quantization noise synchronization is significantly improved if the maximum and minimum of the second and third histogram intervals are tied to the mathematical expectation of biometric data sampling:

where E(x) is the mathematical expectation of each sample, over which the chi-square functional is calculated.

Then the width of the four intervals of the histogram should be calculated by the following formula:

$$\max(\Delta x_2) = \min(\Delta x_3) = E(x) \tag{4}$$

where is the standard deviation of each sample involved in the calculation of chi-square values distribution.

$$\Delta x = \frac{6 \cdot \sigma(x)}{4} \tag{5}$$

The fulfillment of the synchronization conditions (4) and (5) results in the fact that the continuous spectrum of the continuum of the chi-square test states becomes a discrete spectrum. An example of the spectral lines location of chi-square functionals for normal data and uniform initial data is shown in Figure 3. Physics [14, 15] and chemistry [16] study the spectra of hydrogen, oxygen, lithium, sodium and other substances molecules. The spectrum of the output states of the chi-square functional (3) can be considered as some mathematical molecule with four allowed orbitals and 16 electrons on them.

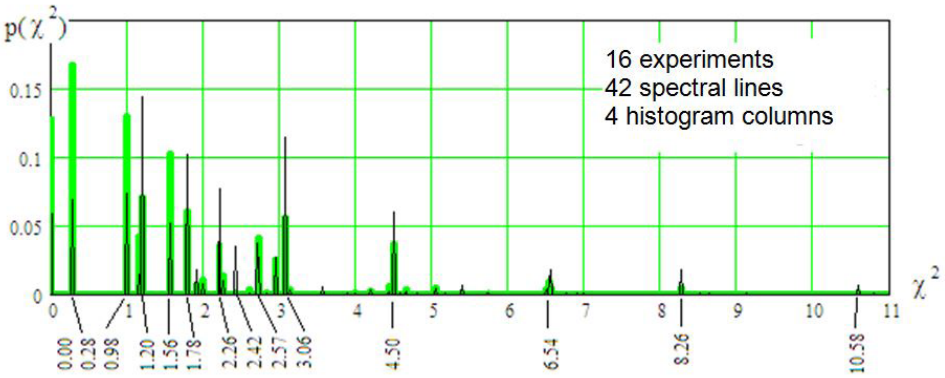

Figure 3: The position of the spectral lines of chi-square functionals (fine lines) for normal data and (thick pale lines) for initial data with uniform distribution

Figure 3 shows that the spectral lines amplitudes of normal and uniform data are different. In fact, one is dealing with two different spectral portraits of uniform and normal data. Different technologies can be used to recognize portraits. In particular, artificial neural networks trained according to GOST R 52633.5 [17] can be used.

The standardized learning algorithm [17] is convenient because it has linear computational complexity and it is possible to estimate the end result of the operation of a standard neural network. In particular, one can estimate the informativity of each spectral component as a module of the logarithm of the amplitudes ratio of identical spectral lines:

It follows from relation (6) that spectral lines with the same amplitude (for example, spectral line No. 15 in Table 1) have zero informativity. On the contrary, the spectral component with the number 13 will have the maximum informativity. It follows from relation (6) that spectral lines with the same amplitude (for example, spectral line No. 15 in Table 1) have zero informativity.

$$I_\Sigma \approx \sum_{i=1}^{20} \left| \log_2 \left\{ \frac{P_{norm}(\chi_i^2)}{P_{ravn}(\chi_i^2)} \right\} \right| \approx \tag{6}$$

24.95 bit

On the contrary, the spectral component with the number 13 will have the maximum informativity.

Table 1: The Spectral Lines Amplitudes of Chi-Square Molecule for Normal and Uniform Input Data

S/N	1	2	3	4	5	6	7	8	9	10
x^2	0.0	0.28	0.98	1.16	1.2	1.56	1.78	1.9	1.98	2.2
Norm.	0.128	0.168	0.130	0.043	0.071	0.103	0.061	0.037	0.011	0.036
Equal to	0.058	0.068	0.074	0.014	0.144	0.053	0.102	0.060	0.003	0.077
S/N	11	12	13	14	15	16	17	18	19	20
x^2	2.26	2.42	2.57	2.72	2.94	3.06	3.12	4.18	4.44	4.5
Norm.	0.015	0.007	0.002	0.042	0.025	0.056	0.003	0.003	0.005	0.056
Equal to	0.005	0.035	0.018	0.018	0.025	0.114	0.005	0.004	0.007	0.114

In the limit, the final informativity of the finite neural network solutions can be estimated as the sum of all particular informativities along the 20 spectral lines of Table 1.

OBTAINING DATA FOR CALCULATING THE SPECTRAL LINES OF THE SAMPLE CHI-SQUARE FUNCTIONAL PORTRAIT FROM 16 EXPERIMENTS

In order to obtain a portrait of the output states spectrum of a chi-square molecule with 16 freedoms, millions of

$$I_{\Sigma} \approx \sum_{i=1}^{20} \left| \log_2 \left\{ \frac{P_{norm}(\chi_i^2)}{P_{ravn}(\chi_i^2)} \right\} \right| \approx \quad (7)$$

24.95 bit

realizations with 16 experiments in each are required. In order to obtain many implementations, the authors take one large sample of 32 experiments. If one randomly selects a small subsample of 16 experiments from a large sample of 32 experiments, then one gets millions of different options:

Further, let us calculate the chi-square functional (3) on each of the millions of subsamples. As a result, a vector of millions of the chi-square molecule states is obtained. Statistical processing of these states will enable data similar to the rows in Table 1.

$$C_{32}^{16} = \frac{32!}{16!(32-16)!} \quad (8)$$

601 000 000

It is extremely important that with a deeper statistical processing of the data in the place of one or two easily computable chi-square functionals (3), the one is compelled to calculate hundreds of millions of such functionals. There is a significant increase in the computational complexity of algorithms for deeper statistical analysis. That is, the algorithms for analyzing the spectral images of small samples may prove to be more powerful than simpler classical algorithms only because they initially require much more computing resources. In Pearson's time, there was no significant computing power. Today the situation is different, one can quite spend several hours of desktop computer work to reduce the requirements for the size of a reliable test sample.

EVALUATION OF THE SAMPLE CHI-SQUARE FUNCTIONAL INFORMATIVITY FROM 32 EXPERIMENTS IN THE CONTINUUM AND DISCRETE VERSIONS

The sample of 32 experiments is small and, therefore, cannot be used for reliable analysis according to standardized requirements [04]. In this connection, it is necessary to involve a numerical experiment built on a million samples from 32 experiments obtained from software generators of pseudo-random numbers with normal and uniform distribution laws. The probability distributions curves of chi-square functionals are shown in Figure 4 for 32 experiments and 4 columns of the histogram.

The informativity of the chi-square test for a sample of 32 experiments is determined by calculating equiprobable errors of the first and second kinds PEE=P1=P2=0.215. Figure 4 shows the equiprobable errors point marked with a dotted line. The sample chi-square functionals in 32 experiments is:

If the estimate (7) and the estimate (9) are compared, a huge gain in the two probabilities ratio is obtained. The probability of an error in classical continuum calculations takes the value of PEE=2-2.2=0.215, transition to the spectral lines analysis reduces the probability of errors up to PEE=2-24.9=0.000000003. There is a 222□4 000 000-fold gain on reduction of errors of the first and the second kind. This is precisely the stimulus to force researchers to move from a relatively simple classical continuum statistical processing to a more complex image-processing of spectral portraits of small samples.

$$I = -\log_2(P_{EE}) = 2.21 \text{ bit} \quad (9)$$

Moreover, it can be shown that the gain can be even greater if one goes from using histograms with 6 columns instead of histograms with 4 columns. In this case, the number of significant spectral states approximately doubles, which is shown in Figure 5.

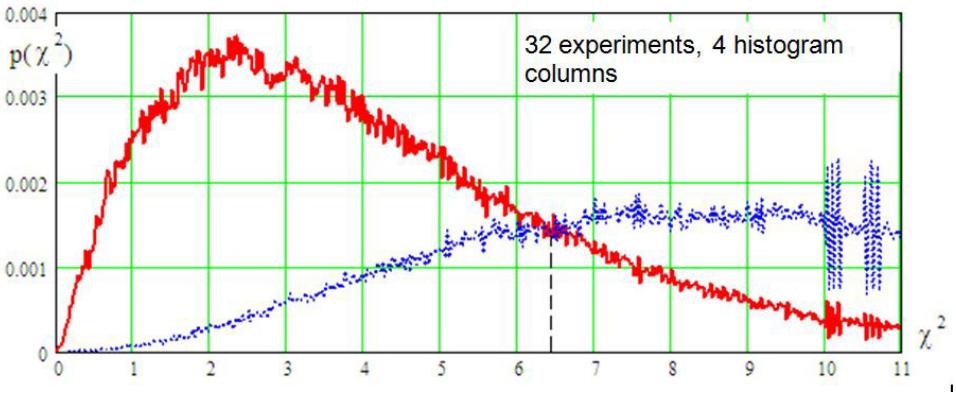

Figure 4: The densities of distribution values of the chi-square functionals for the normal (solid line) and uniform (dashed line) laws

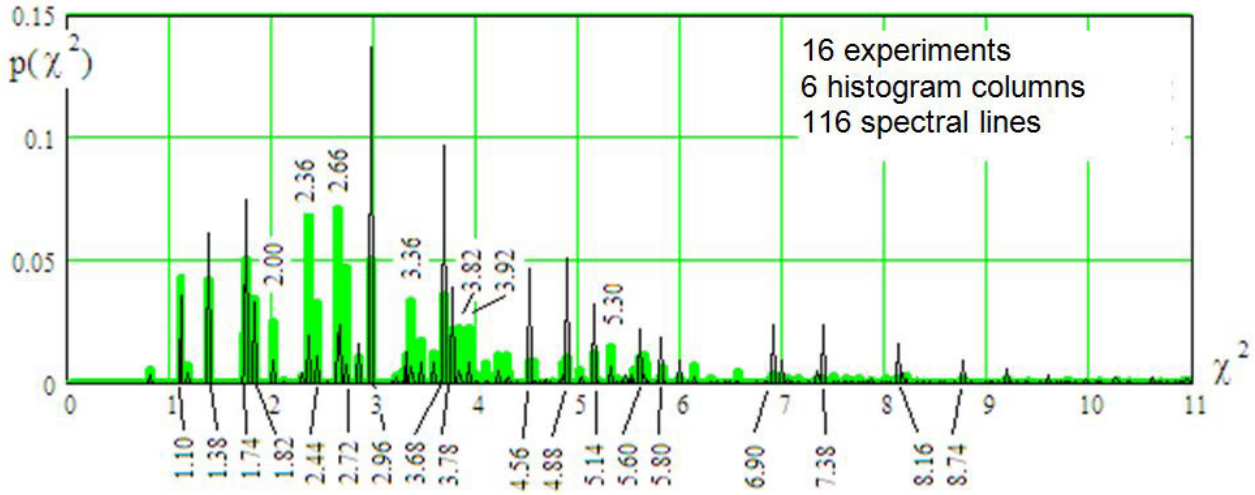

Figure 5: The discrete spectrum of chi-square test conditions, which has almost doubled the number of spectral lines with an increase in columns in the histogram

As a result, estimate of errors of the first and second kind probability in the transition from histograms with 4 columns to histograms with 6 columns should decrease significantly with PEE=2-24.9 to PEE=2-49.8 (the exponent value should at least double).

CONCLUSION

It should be noted that the gain estimate from the continuous probability densities analysis to the discrete spectral lines analysis is an upper-bound estimate. Real gain will be always lower. This is due to the fact that the spectral lines of the chi-square molecule are correlated (linked together [18]). Correlation links between the data arise because a set of samples in the size of 16 experiments are obtained from one sample in 32 experiments. However, even this circumstance should not stop the researchers of such statistics. Even if the theoretically possible gain of 4,000,000 times is reduced to a technically realizable gain of 40 times, the programming costs and additional power consumed by the computers will be justified when processing small amounts of expensive biometric data.

Gain from a deeper statistical processing of the second level should always be present. This gain can be estimated more reliably only after the implementation of the corresponding software product. As a matter of fact, in the offing, there is a technical opportunity to exchange the complexity of the software product and the time for solving problems on the computer on the reliability of the statistical decisions it makes. Even today, when calculating the standard deviation and the correlation coefficients, it is possible to reduce the requirements for the test sample volume from 1.2 to 3 times. In this article, the authors tried to demonstrate that the potential for further reducing the requirements for the test sample size is much larger.

REFERENCES

1. Volchikhin, V.I., Ivanov, A.I., Funtikov, V.A. (2005). Fast learning algorithms for neural network mechanisms of biometric-cryptographic information protection. Publishing House of the Penza State University, Penza.
2. Malygin, A.Yu., Volchikhin, V.I., Ivanov, A.I., Funtikov, V.A. (2006). Fast testing algorithms for neural network mechanisms of biometric-cryptographic information protection. Publishing House of the Penza State University, Penza.
3. Yazov, Yu.K. (2012). Neural network protection of personal biometric data. Radio Engineering, Moscow.
4. Federal Agency on Technical Regulating and Metrology. (2001). R 50.1.037-2002. Recommendations on standardization. Applied statistics. Rules for verifying the agreement between the experimental and the theoretical distributions. Part I. x2 type criteria.
5. Serikova, N.I., Ivanov, A.I., Kachalin, S.V. (2014). Biometric stats: smoothing histograms based on small training sample. Scientific Journal of Science and Technology, vol. 3, no. 55, 146-150.
6. Serikova, N.I. (2015). Assessment of the likelihood of the normal distribution hypothesis by the Gini criterion for smoothed histograms constructed on small test samples. Questions of Radio Electronics, vol. 1, 85-94.
7. Ivanov, A.I., Akhmetov, B.B., Serikova, Yu.I. (2016). Strengthening the power of the chi-square test with tenfold increase in freedoms of statistical computations on small test samples. Reliability and Quality of Complex Systems, vol. 4, no. 16, 121-127. DOI: 10.21685/2307-4205-2016-4-17

8. Akhmetov, B.B., Ivanov, A.I. (2016). Multidimensional statistics of essentially dependent biometric data generated by neural network emulators of quadratic forms. LEM, Almaty.
9. Akhmetov, B.B., Ivanov, A.I., Serikova, N.I., Funtikova, Yu.V. (2015). The discrete nature of the chi-square test distribution for small test samples. Bulletin of the National Academy of Sciences of the Republic of Kazakhstan, vol. 1, no. 353, 17-25.
10. Kulagin, V.P., Ivanov, A.I., Gazin, A.I., Akhmetov, B.B. (2016). Cyclic continuum-quantum computing: Strengthening the Chi-Square test power on small samples. Analytics, vol. 30, no. 5, 22-29.
11. Volchikhin, V.I., Ivanov, A.I., Serikov, A.V., Serikova, Yu.I. (2017). Quantum superposition of the state discrete spectrum of mathematical correlation molecule for small samples of biometric data. Mordovia University Bulletin, vol. 27, no. 2, 224-238. DOI: 10.15507/0236-2910.027.201702.224-238
12. Volchikhin, V.I., Ivanov, A.I. (2017). Neural Network Molecule: a Solution of the Inverse Biometry Problem through Software Support of Quantum Superposition on Outputs of the Network of Artificial Neurons. Mordovia University Bulletin, vol. 27, no. 4, 518-529. DOI: 10.15507/0236-2910.027.201704.518-529
13. Volchikhin, V.I., Ivanov, A.I., Gazin, A.I., Bannih, A.G. (2017). Conditions of obtaining the discrete kurtosis spectrum of statistical distributions of biometric data for small samples. Journal of Computational and Engineering Mathematics, vol. 4, no. 4, 53-63. DOI: 10.14529/jcem170405
14. Nilson, M., Chang, I. (2006). Quantum calculations and quantum information. Mir Publishers, Moscow.
15. Filippov, V.V., Mitsuk, S.V. (2017). Modelling magnetoresistance effect in limited anisotropic semiconductors. Chinese Physics Letters, vol. 34, no. 7, 077201. DOI: 10.1088/0256-307X/34/7/077201
16. Stepanov, N.F. (2001). Quantum mechanics and quantum chemistry. Mir Publishers, Moscow.
17. State Standard R 52633.5-2011. (2011). Data Protection. Information Protection Technique. Automatic Learning Neural Network Converters Biometry-Code Access. Standartinform, Moscow.
18. Ivanov, A.I. (2016). Multidimensional neural network processing of biometric data with software reproduction of quantum superposition effects. Penza Scientific and Research Electronic Technical Institute, Penza.

Paper submitted: 27.07.2018.
Paper accepted: 14.08.2018.