Possibility of Decrease in a Level of Data Correlation During Processing Small Samples Using Neural Networks by Generating New Statistic Tests

A I Ivanov$^1$, A G Bannykh$^2$, P S Lozhnikov$^3$, A E Sulavko$^3$, and D P Inivatov$^3$

$^1$Penza Research Institute of Electrical Engineering, Penza, Russia
$^2$Penza State University, Penza, Russia
$^3$Omsk State Technical University, 11 Mira ave., Omsk, 644050, Russia

Abstract. Statistic tests created in the 20th century may be mapped with some equivalent artificial neurons. As a result, a network of dozens of artificial neurons that combines dozens of known statistic tests may be used for the validation of the normality hypothesis. The quality of solutions made by a neural network depends on a number of used neurons (tests). This tendency gives rise to the task of creating new statistic tests (neurons) that first and foremost require low correlation of their decision with known tests. The paper presents a forecast of attainable confidential probabilities for the validation of the normality hypothesis for a small sample of 21 tests in a network consisting of 21 artificial neurons, where each one is mapped with one traditional statistic test. When new tests are used (that should be created in the 21st century), the correlation of data is expected to lower by far, which should allow an approximately 10-fold decrease in a number of error probabilities.

1. Statement of a problem

An insufficient size of a sample (training, test, etc.) is a big challenge in machine learning and mathematical statistics [1, 2]. In particular, this problem is discernible in the validation of the hypothesis of the random distribution law. It is not possible to validate the normality for a small-sized sample using, for example, the chi-squared test. If the standardized guides (GOST R 50.1.037-2002) of the chi-squared test are applied, the confidential probability of 0.99 is attained when samples of 140 and more tests are used. When the standardized automatic neural network training is applied for converting a human biometric image to a cryptographic key code (GOST R 52633.5-2011), a sample of 21 tests is enough for obtaining congruent confidence in the decisions. This raises a question about the adequateness of the results that were obtained with modern practices (GOST R 52633.5-2011) and the traditional approaches (GOST R 50.1.037-2002).

The reason for this discrepancy is the usage of a new tool in the 21st century – self-trained large artificial neural networks. Large neural networks working on modern computers can perform very complicated computations. This resource was not available for researches in the 20th century.

Another question is a limit of accuracy for neural network generalization of traditional statistical calculations. The problem of “the curse of dimensionality” is widely known. When the task gets more complicated, the computational stability falls and errors accumulate. However, there is always a technically achievable range when calculations output a reliable result.

In this paper, we will try to demonstrate that positive experience of neural network biometry may be reasonably used for creating neural network generalizations of a large number of statistical tests that were developed last century. Furthermore, dozens of existing statistic tests may be supplemented by dozens of new statistic tests that will be created in the 21st century specific requirements of increasing the efficiency of neural network generalization.
2. Software neuron equivalent to chi-squared test

Pearson’s chi-squared test for a sample of 21 tests is easy to imagine as a certain software squared neuron that sorts initial data, estimates the number of tests in 5 histogram bins. Later the summation function accumulates the discrepancies for the predicted and experimental values. At the summation function output, the quantization is performed. The functional connections of Pearson’s squared neuron with 5 inputs are given in Table 1. The training of this neuron is done in the framework of the differentiation of the normal and uniform distribution laws providing type I and type II errors are of the same probability.

Table 1. Chi-squared neuron for the sample of 21 tests

| \( x \leftarrow \text{sort}(x) \) | \( \Delta \leftarrow \frac{x_{20} - x_0}{5} \) | \( \bar{x}_i \leftarrow x_0 + \Delta \cdot i; \ i = 0, 1, ..., 5 \) |
| --- | --- | --- |
| \( \chi^2 \leftarrow 21 \cdot \sum_{i=0}^{4} \left( \frac{n_i}{21} - (P(\bar{x}_{i+1}) - P(\bar{x}_i)) \right)^2 \) | where \( \Delta \) is a histogram bin width, \( n_i \) is the number of samplings got into the \( i^{\text{th}} \) histogram bin, \( P(\bar{x}_i) \) is predicted probability for the normal data distribution, \( P_1 \) is type I error probability, \( P_2 \) is type II error probability. |

\[
\begin{align*}
\chi^2 & \sim \chi^2(21) \\
z(\chi^2) = & \text{"0" if } \chi^2 \leq 7.5 \\
z(\chi^2) = & \text{"1" if } \chi^2 > 7.5 \\
P_1 & \approx P_2 \approx P_{EE} \approx 0.292
\end{align*}
\]

The software Pearson’s artificial neuron (see Table 1) requires about a million of small samples of 21 tests. For data with the normal distribution law and the uniform distribution law, two densities of value distributions are shown in Fig. 1.

As Fig.1 shows, when the quantization threshold is 7.5 type I and type II error probabilities are close, and the confidential probability of the correct data separation is 0.708, which is inappropriate for practical purposes. It is clear that in a similar way other equivalent software neurons may be built for other statistic tests. They all will be described by functional connections similar to the connections from Table 1, but equally possible values of the type I and type II errors will differ.

Table 2 shows the values of equally possible errors PEE for 8 different statistic tests (neurons).

Table 2. Values of error probabilities for validation tests of statistical hypothesis using samples of 21 tests

| #  | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  |
|----|----|----|----|----|----|----|----|----|
where
\[ \chi^2 \] is chi-squared test [3, 4] (1900);
\[ ad^2 \] is Anderson-Darling test [4] (1952);
\[ adL \] is a logarithmical form of Anderson-Darling test [3, 4] (1952);
\[ sg \] is the geometrical mean test [4] (2014);
\[ sg_d \] is a differential and integral variant of the geometrical mean [4] (2016);
\[ \omega^2 \] is the Cramér–von Mises criterion [3, 4, 5] (1928);
\[ \omega^2_c \] is the Smirnov–Cramér–von Mises test [3, 4] (1936);
\[ su^2 \] is the Shapiro–Wilk test [3] (1965).

It is obvious that 8 statistic tests instead of one can be easily performed on modern computers. Moreover, the neural network implementation of this technical solution will result in approximately 8-fold complication of computation and output «00000000» as a code state when all tests (neurons) decide in favor of the normal value distribution. If all neurons decide in favor of uniform value distribution, we get «11111111» as an output code.

In practice, the output code of the neural network does not often have equal states. In such cases, the decision is made based on the majority of the observed states. That means all codes with the majority of “0” states are treated as a decision in favor of the normal distribution of values for a sample of 21 tests.

If we are guided by a common practice of utter disregard of correlation relations between neuron output values, «00000000» states should be treated as a decrease in values of equally possible errors to the quantity:

\[ P_{EE,8} \approx \prod_{i=1}^{8} P_{EE,i} \approx 0.0001 \]  

Unfortunately, this primitive approach does not work. The meaning (1) is too optimistic; it is not allowed to disregard the correlation relations of the neuron output states.

3. Taking into consideration the correlation relations of neuron output states

As correlation relations should be taken into account, the algorithm from the standard GOST P 52633.3-2011 is used in neural networks. The standard algorithm bypasses the problem of the exponential growth of the computing complexity of the entropy calculation using the Shannon formula for the codes with the length of 2596 bits. To decrease the computing complexity to a linear one, it is necessary to move from the analysis of a typical code to an analysis of the Hamming distances.

Yet another important consequence of such transformations is the normalization of the Hamming distance distributions for the codes of 32 bits and more. But the most important is that the normalization of the data gives a possibility to symmetrize the task. The symmetrization appears in the following way:

\[
\begin{bmatrix}
1 & r_1 & r_2 & \cdots & r_n \\
1 & r_1 & r_{n+1} & \cdots & r_{2n-2} \\
1 & r_{n+1} & 1 & \cdots & r_{3n-3} \\
1 & r_{2n-2} & r_{3n-3} & \cdots & 1 \\
\end{bmatrix}
\rightarrow
\begin{bmatrix}
1 & \tilde{r} & \tilde{r} & \cdots & \tilde{r} \\
\tilde{r} & 1 & \tilde{r} & \cdots & \tilde{r} \\
\tilde{r} & \tilde{r} & 1 & \cdots & \tilde{r} \\
\tilde{r} & \tilde{r} & \tilde{r} & \cdots & 1 \\
\end{bmatrix}
\]  

Instead of a real correlation matrix of real neuron data, an equivalent correlation matrix [6] of equally correlated neurons appears. The condition of the accuracy of the symmetrization procedure (2)
involves the concordance of equal probabilities for type I and type II errors of the first asymmetric neural network and the second symmetric neural network.

The easiest way of computing the rates of equal correlation is module averaging of all correlation coefficients of the initial asymmetric matrix. This method works for a task with a dimension of 32 and more very well. However, it is difficult to validate the equivalence of the conversion by a direct numerical experiment using ordinary computing machines.

The correlation coefficients of output states of the corresponding neurons for this case of 8 tests generalized by a neural network are given in Table 3.

Table 3. Correlation coefficients between pairs of statistic tests under consideration

|        | $\chi^2$ | ad$^2$ | adL  | sg   | sg$_d$ | $\omega^2$ | $\omega^2_c$ | su$^2$ |
|--------|---------|--------|------|------|--------|-----------|-------------|------|
| $\chi^2$ | 1       | 0.423  | 0.672| 0.037| -0.042 | 0.559     | 0.401       | -0.726|
| ad$^2$  | 0.423   | 1      | 0.644| 0.018| -0.145 | 0.226     | 0.393       | -0.113|
| adL    | 0.672   | 0.644  | 1    | 0.056| 0.209  | 0.827     | 0.832       | -0.917|
| sg     | 0.037   | 0.018  | 0.056| 1    | 0.132  | 0.414     | 0.402       | -0.212|
| sg$_d$ | -0.042  | -0.145 | 0.209| 0.132| 1      | -0.242    | -0.142      | -0.041|
| $\omega^2$ | 0.559 | 0.226  | 0.827| 0.414| -0.242 | 1         | 0.885       | -0.667|
| $\omega^2_c$ | 0.401 | 0.393  | 0.832| 0.402| -0.142 | 0.885     | 1           | -0.764|
| su$^2$  | -0.726  | -0.113 | -0.917| -0.212| -0.041 | -0.667    | -0.764      | 1     |

Averaging of correlation coefficient modules from the Table 3 provides an estimation of equal correlation $\tilde{r} \approx 0.398$. The geometric mean of equally possible errors of the tests (neurons) from the Table 3 is:

$$\tilde{P}_{EE} = \sqrt[8]{\prod_{i=1}^{8} P_{EE,i}} \approx 0.316$$

(3)

Having these two parameters and applying the simulation modeling, we can obtain a graph of continuous decrease of values of equal probabilities for type I and type II errors as the number of neurons increases. The graph $P_{EE,n}$ appears to be complicated enough, but the situation changes if logarithmical coordinates use. Figure 2 shows two nearly linear graphs for different values of equal correlation coefficients.

Fig. 2 shows that the symmetrized network of 8 artificial neurons must provide $P_{EE} \approx 0.02$. This prediction is approximately 100 times more reasonable than the prediction (1) that computationally proves a necessity to take into account real correlation relations of the neuron output states. The complete disregard of the correlation relations is a bad practice of the previous century fuelled by a subconscious focus on the economy of the computing resources.
4. Shift of the paradigm of creating new statistic tests

Pearson created his chi-squared test in 1900 and it became a motivation for hundreds of mathematicians who created about 200 statistic tests [3] during the 20th century. The created statistic tests were optimized based on their power in comparison with other tests. In fact, only those tests that provided lower values of $P_{EE}$ under one or another condition were published. If the power of the test were lower there was no sense to publish it. That means the statistical magazines of the 20th century recorded only relatively power statistic tests. A majority of relatively weak statistic tests were investigated for sure but not published.

These relatively powerful statistic tests are ranged with respect to each other. In particular, A.I.Kobzar [3] provides a table consisting of 4 variants of their ranging for the task of validating the normality hypothesis for the 21st statistic test. If we generalize all 21 statistic tests by a network of 21 neurons, the error probability value should decrease up to the value $P_{EE} \approx 0.0045$ (the lower line in Fig. 2). This probability is quite enough for a majority of practical purposes. However, for traditional statistic tests, this error probability is unachievable. The problem is that the traditional statistic tests, as a rule, are highly correlated.

Table 3 shows correlation coefficients of 6 traditional tests and 2 relatively new tests (sg, sgd). If we disregard new tests and average only correlation coefficient modules for traditional tests the estimation of the equal correlation increases up to the value $\tilde{r} \approx 0.603$ (the second line with gentler gradient). The type I and type II error probabilities decrease slower with a rise of a number of generalized statistic tests. As a result, the expected value of error probabilities for the 21st statistic test reach runs to $P_{EE} \approx 0.022$ only. That means the parallel computation of the 21 traditional statistic tests in a neural network must work worse than 8 neurons described above. A nearly 3-fold increase in the number of tests does not provide significant improvement in the quality of decisions made because of an increase in the correlation (correlation connectivity of neurons).
It turns out that it is not enough to aim at a growth of their relative capacity while creating new statistic tests in the 21st century. It is more important to create new tests that are low correlated with a group of known statistic tests.

5. An example of generating a new statistic test and validating a level of its correlation connectivity with traditional tests

Let us consider a traditional kurtosis test or the normalized fourth moment of the distribution. The implementation of a neural network equivalent for this statistic test is shown in Table 4.

Table 4. Kurtosis neuron for a sample of 21 tests

| Condition | Formula |
|-----------|---------|
| if $\mu_4 \leq 2.13$ | $z(\mu_4) = 1$ |
| if $\mu_4 > 2.13$ | $z(\mu_4) = 0$ |
| $P_1 \approx P_2 \approx P_{EE} \approx 0.212$ |

where $E(x)$ is a mathematical expectation of the sample,

\[ \sigma(x) \text{ is the standard deviation of a small sample,} \]

\[ corr(\chi^2, \mu_4) = 0.07, \]  \[ corr(\omega^2, \mu_4) = 0.145, \]

\[ corr(\omega^2_c, \mu_4) = 0.164, \]  \[ corr(ad^2, \mu_4) = 0.486, \]

\[ corr(adL, \mu_4) = 0.223, \]  \[ corr(sg, \mu_4) = 0.651, \]

The distribution of values of this neuron-test while impacting it by small samples with normal and uniform distributions are shown in Fig. 3.

We modify the kurtosis test by dividing by the density of the normal distribution law. This modification of the neuron-test is described by functional relations shown in Table 5.

Table 5. Kurtosis neuron for a sample of 21 tests

| Condition | Formula |
|-----------|---------|
| if $\mu_4 \leq 2.13$ | $z(\mu_4) = 1$ |
| if $\mu_4 > 2.13$ | $z(\mu_4) = 0$ |
| $P_1 \approx P_2 \approx P_{EE} \approx 0.212$ |

where $p(x)$ is normal density of distribution of values of a small sample,

\[ corr(\mu_4, \tilde{\mu}_4) = 0.619, \]

\[ corr(\chi^2, \tilde{\mu}_4) = 0.05, \]  \[ corr(\omega^2, \tilde{\mu}_4) = 0.165, \]

\[ corr(\omega^2_c, \tilde{\mu}_4) = 0.279, \]  \[ corr(ad^2, \tilde{\mu}_4) = 0.612, \]
The responses of the summation function of the modified neuron-test are shown in Fig. 4.

\[ \bar{\mu}_4 \leftarrow \frac{1}{21} \sum_{i=1}^{21} \left( (E(x) - x_i)^4 \right) \]

\[ z(\bar{\mu}_4) \leftarrow "1" \text{ if } \bar{\mu}_4 \leq 23.7 \]

\[ z(\bar{\mu}_4) \leftarrow "0" \text{ if } \bar{\mu}_4 > 23.7 \]

\[ P_1 \approx P_2 \approx P_{EE} \approx 0.198 \]

The new modification of the initial test would have absolute priority if the only purpose was an increase in the test capacity. If we calculate the correlation connection between the new test with the tests from Table 3, the single-valuedness disappears. It is worth adding a single traditional test of the forth statistic moment to the tests from the Table 3 because it provides a larger decrease in mean correlation by a group as the capacity decreases at a very slight degree. It is more useful to add both considered tests. The modeling situation clearly shows that a network of 9 neurons-tests always works worse than a network of 10 neurons-tests. The reason is that both adding tests have a low correlation connection with the majority of tests from Table 3.

6. Formalization of some simple rules of generating new statistic tests by modifying known ones

6.1 Multiplication (devision) of the kernel of integral by the probability functions and their derivatives

Some statistic tests have analytical descriptions. In particular, one of them is a traditional criterion by Cramér–von Mises created in 1928:

\[ \omega^2 = \int_{-\infty}^{\infty} \left( P(u) - \tilde{P}(u) \right)^2 \cdot du \]  

(4)

where \( P(u) \) is a theoretical probability function, \( \tilde{P}(u) \) is an experimental probability function.

In 8 years, in 1936 Smirnov modified the Cramér–von Mises criterion:
\[ \omega_c^2 = \int_{-\infty}^{\infty} \left[ P(u) - \tilde{P}(u) \right]^2 \cdot dP(u) \]  

(5)

It is obvious that the criterion is modified by multiplying the subintegral kernel of the criterion by the probability density function:

\[ \omega_c^2 = \int_{-\infty}^{\infty} \left[ P(u) - \tilde{P}(u) \right]^2 \cdot p(u) \cdot du \]  

(6)

where \( p(u) = \frac{dP(u)}{du} \) is differential density of the probability of the theoretical distribution.

It should be noted that in 1936 Smirnov suggested his modification, transforming the square subintegral space of data accumulation. The non-linear distortion is done by multiplication of the kernel of the integral by the differential density of the theoretical distribution probability.

In the previous paragraph, the forth statistic moment test was modified in the same way. The only distinction is that the modification of the non-linear space of data accumulation (summation) is done by dividing by the function of the probability density for the theoretical distribution instead of multiplying.

6.2. Substitution of the integral probability function for the differential density of value distribution

Nominally any non-linear transformation of the initial square difference space of accumulation must lead to a certain non-correlated component in a response of a test-descendant in regard to a test-parent. In this relation the criterion (6) may be modified by substituting of the distribution density by the integral probability function:

\[ \omega_{CM}^2 = \int_{-\infty}^{\infty} \left[ P(u) - \tilde{P}(u) \right]^2 \cdot P(u) \cdot du \]  

(7)

In regard to such modification, additional computational researches are required that demonstrate to what extent the new test (7) will be capable and correlated with other known tests.

However, a possibility of modification for statistic tests by simple substitution of the integral functions of probability for their differential probability densities is validated in a numerical way. In particular, it has been done using differential variants of the Cramér–von Mises criterion family [4, 5] as an example:

\[ \omega_d^2 = \int_{-\infty}^{\infty} \left[ p(u) - \tilde{p}(u) \right]^2 \cdot du \]  

(8)

\[ \omega_{Cd}^2 = \int_{-\infty}^{\infty} \left[ p(u) - \tilde{p}(u) \right]^2 p(u) \cdot du \]  

(9)

Such substitution of probability functions for their differential equivalent gives responses of modified neurons-tests that are not significantly correlated. A coefficient of average correlation of such transformations regarding a parent test ranges from 0.5 to 0.7, which is accessible for neural network generalizations.

6.3. Shift of a space type of accumulating (enriching) data by an artificial neuron or a new statistic test

It should be noted that the Cramér–von Mises criterion created in 1928 (4) accumulates (enriches) the data in a square difference space. In this context, the Gini coefficient developed in 1941 may be taken as a modification of the Cramér–von Mises criterion, which was obtained by the distortion of the non-linear space by square-rooting:
\[
D = \int_{-\infty}^{+\infty} (P(u) - \bar{P}(u))^2 \cdot du = \int_{-\infty}^{+\infty} (P(u) - \bar{P}(u)) \cdot dP(u)
\]

(10)

In this case, the Frosini criterion created in 1978 may be taken as a modification, which was used in Smirnov for the Cramér–von Mises criterion in 1936 but without square-rooting:

\[
Fr = \int_{-\infty}^{+\infty} (P(u) - \bar{P}(u))^2 \cdot dP(u) = \int_{-\infty}^{+\infty} (P(u) - \bar{P}(u)) \cdot p(u) \cdot du
\]

(11)

Unfortunately, the substitution of square for the module computation provides a slight effect on a decrease in correlation. A far more effect is obtained if the probability functions substitute for their differential equivalent at the same time:

\[
D_d = \int_{-\infty}^{+\infty} (p(u) - \bar{p}(u))^2 \cdot du = \int_{-\infty}^{+\infty} (p(u) - \bar{p}(u)) \cdot dP(u)
\]

(12)

The correlation relations could be decreased much more if instead of raising to the square we compute the geometric mean for the probability characteristics considered [4, 5]:

\[
sg = \int_{-\infty}^{+\infty} \sqrt{P(u) \cdot (1 - \bar{P}(u))} \cdot du
\]

(13)

The researches carried out show that the integro-differential variant of the geometric mean criterion is one of the most efficient and less correlated with other known tests at the same time:

\[
sg_d = \int_{-\infty}^{+\infty} \sqrt{p(u) \cdot \bar{P}(u)} \cdot du
\]

(14)

In the past century, the creation of any of the traditional statistic tests was a significant event as their authors relied on the knowledge of asymptotic statistic relations. Each of the traditional statistic tests of the past century was created nearly in an analytical way “with the point of a pen”. Today the situation has changed. Any engineer or a software developer can modify known tests described in this paper.

7. Conclusion

The comparative ease of generating new statistic tests makes us hope that the list of 21 most commonly used tests for normality hypothesis validation may be doubled at least. Moreover, there are no reasons against keeping a generally low level of correlation connection of data at \( \tilde{r} \approx 0.4 \) for 42 neurons based on the corresponding tests. In this situation, we can expect a decrease in error probability up to 0.0017 (see Fig. 2) in the nearest future. The more pessimistic approach is based on the idea that the correlation connection for 42 neurons-tests will be the same as for known tests (\( \tilde{r} \approx 0.6 \)). In this case, the error probability will be 0.013 that is quite acceptable for practical use. In both cases, while new statistic tests are generated, the average group correlation connection is more important than their relative capacity. Heading for the neural network generalizations of a large number of statistic tests, one should aim at the decrease in correlations of their decisions.

Neural network generalization of a large number of comparatively weak and comparatively dependent statistic tests is expected to strengthen trust in the statistic validation performed with the use of small samples. And an increase in a computing complexity \( u \ p \) to 42 times is not a meaningful factor for modern computers.

References:
[1] Vabalas A, Gowen E, Poliakoff E and Casson AJ (2019) Machine learning algorithm validation with a limited sample size PLoS ONE 14(11): e0224365

[2] Ivanov A I, Lozhnikov P S and Sulavko A E 2017 Evaluation of signature verification reliability based on artificial neural networks, Bayesian multivariate functional and quadratic forms (Samara: Computer Optics) №5 pp 765-774

[3] Kobzar A I (2006) Applied Mathematical Statistics. A Handbook for Engineers and Researchers (Moscow: Fizmatlit) p 816 (in Russian)

[4] Ivanov A I 2016 Multidimensional neural network processing of biometric data with software reproduction of quantum superposition effects Monograph (Penza: PNIEI) p 133 (in Russian)

[5] Sulavko A E, Zhumazhanova S S and Fofanov G A 2018 Perspective Neural Network Algorithms for Dynamic Biometric Pattern Recognition in the Space of Interdependent Features XII Int. IEEE Scientific and Technical Conference «Dynamics of Systems, Mechanisms and Machines» (Omsk: Omsk State Technical University) pp 1-12

[6] Ivanov A I, Lozhnikov P S and Serikova Y I 2016 Reducing the Size of a Sample Sufficient for Learning Due to the Symmetrization of Correlation Relationships Between Biometric Data Cybernetics and Systems Analysis (Electronic Materials vol 52) № 3 pp 379–385