ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Assessing the Level of Uncertainty of small samples of Multidimensional Biological and Biometric Data

Bakhytzhan Akhmetov [1*], Alexander Ivanov 2, Elena Malyghina [3]
Sergey Kachalin [3], Natalya Serikova [3]
[1] Kazakh National Technical University named after K.I.Satpayev, Kazakhstan, Almaty
[2] Penza Scientific-Research Electrotechnical Institute, Russia, Penza
[3] Penza State University, Russia, Penza
b_akhmetov@ntu.kz

Abstract

It is shown that small selections of biometric data of the person have essential uncertainty which needs to be considered at multidimensional statistical estimates. Within a hypothesis of normal distribution of the bio-these the table of errors of calculation of a population mean and a mean square deviation as functions of number of examples in test selection is constructed. The table can be used for an assessment of an error of the received results of biometric measurements, and also for a reasonable variation by data in attempts of artificial increase in the amount of small selections by crossing of examples parents and receiving examples descendants. It is claimed that at the small volumes of basic data their uncertainty is a source of mutations, more powerful, than recommended by the standard.

Keywords: the apparatus of artificial neural networks, statistical procedures for the processing of biometric data, the small sample, the test sample

Introduction

In systems of recognition of biometric images of the person about 512 signs are used. The most effective for the solution of a task today is the device of artificial neural networks [1-3]. Efficiency of application of artificial neural networks in the analysis of biological and biometric data is caused by that they are capable to be trained, being arranged under features of object of research. The more dimension of an artificial neural network used at data processing, the more effective appears processing of biometric data. Allegedly that the new techniques of neuronetwork data processing created during the period since 2000 to the present, will find application and for processing of experimentally obtained biological data.

The first of the main a biometrics problem for today is training of big neural networks on the small training selections containing parameters from 11 to 21 examples of a recognizable image. The second of the main problems is a problem of reliable testing of the received results of neuronetwork processing on test selections of limited volume from 21 to 42 examples. The nature of a set of biometric signs and biological signs is identical. In both cases we have strongly correlated data: growth, weight, volume of fragments of a body, distances between the control chutes, results of analyses of biological tests, behavioural characteristics.

Biologists face the same problems, as developers of technical appendices of biometric identification of the identity of the person. Allegedly that a number of the procedures of neuro- and statistical data processing standardized for biometrics can be transferred to biology, without any modifications. In biology, medicine and biometrics problem of insufficient volume of an available statistical material is everywhere. Doctors always don't have rather big selection of "patients" by a rare disease against unlimited selection "healthy", biologists have no rather big selection rare and therefore the most interesting individuals in some population. In biometrics, as a rule, there are no examples of an image "Own" against rather representative selection "Strangers".

One of solutions of this task is repetition in selection of data or withdrawal of data from the selection, carried out, so-called, a butstrap-method

[4]. One more standard reception for biometrics is crossing of examples parents and the receiving from them of examples streams which are carried out in accordance with GOST P 52633.2-2010 [5].

This article is devoted to specifications of statistical procedures of processing of biometric data which, according to her authors, are applicable both in medicine and in biology. Statements, assumptions and the used methods of statistical processing are actual in all three directions of researches.

Basic assumptions

One of the main assumptions is that data quite a lot, tens or even hundred biological (biometric) parameters are controlled. All biometric parameters are significantly connected among themselves; coefficients of pair correlation existing between controllable parameters have density of distribution of values the example of one of which is displayed in Figure 1.

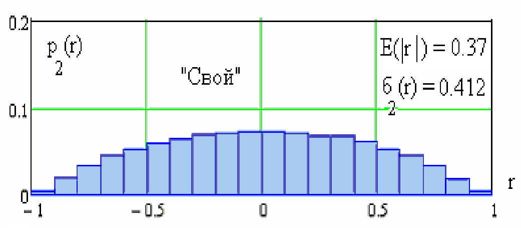

Fig. 1 an Example of the density distribution of the values of the coefficients of pair correlations between biometric data of the image of "Own".

From figure 1 it is visible that the population mean of modules of coefficients of correlation makes 0.37. Besides, we will use a hypothesis of a normality of distribution of multidimensional dependent data, that is, they have to be described by the volume of some multidimensional hyper ellipse. It is impossible to display a hyper ellipse on flat paper, in this regard we will consider only a two-dimensional case. The example of some two-dimensional section of a hyper ellipse is given in figure 2.

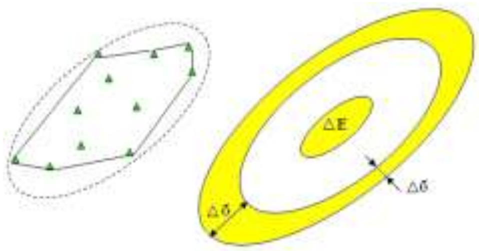

Fig. 2 One of the sections of hyper ellipses in its discrete and analytical view

In the left part of figure 2 eleven projections of 11 examples of training selection to considered hyperplane. It is obvious that attempt to delineate examples in training selection will give a considerable error of approach of real data. There is the error of the sampling caused by small number of examples in presented selection. Obviously, desire to leave from errors of sampling due to integrated procedures of calculation of a vector of population means of biometric parameters $E(\nu_i)$ and their vector of their mean square deviations of $\sigma(\nu_i)$. Knowing a vector of these statistical moments, it is possible to calculate coefficients of pair correlation between the biometric parameters $r_{i,j}$ necessary to us. In turn from correlation coefficient we can pass to a ratio of a big half shaft – *a* and a small half shaft – *b* of the ellipse which has appeared in the considered section:

$$\frac{1+r(\nu_i,\nu_j)}{1-r(\nu_i,\nu_j)}=\frac{a}{b} \tag{1}$$

It turns out that having information about the mathematical expectation of two bio-parameters, standard deviations and correlations between them, we can build the appropriate ellipse two-dimensional distribution of "Own" image data (the dotted line in the left part of figure 2). Data outside of the ellipsoid will correspond to the data of images of "Stranger", and the data inside an ellipsoid will match the image of "Own".

If we need to describe the decision rule of multidimensional statistical analysis, we have to use the corresponding quadratic form:

$$D^2=(E(\bar{\nu})-\bar{\nu})^T\cdot[\rho]^{-1}\cdot(E(\bar{\nu})-\bar{\nu}) \tag{2}$$

where $[\rho]^{-1}$ is the inverse matrix of the coefficients of covariance, each of its elements can be defined through the corresponding correlation coefficient:

$$\rho(\nu_i,\nu_j)=\sigma(\nu_i)\cdot\sigma(\nu_j)\cdot r(\nu_i,\nu_j) \qquad (3)$$

Quadratic form (2) has important theoretical significance, however, for practical calculations it is less applicable. The problem is bad conditionality for the treatment of covariance matrices. On small samples from 20 examples of bio data younger statistical moments (mean value, standard deviation, correlation coefficients) are estimated with large errors. For this reason instead of high-dimensional classical matrix transformations (2), are compelled to use the big artificial neural networks which training is less sensitive to mistakes because of the insufficient volume of basic data. The purpose of this work is creation of tables, an assessment of value of errors of calculation of the younger statistical moments for different number of data (freedom degrees) in small training selection or small test selection.

Assessment of mistakes, calculations of the younger statistical moments

Obviously, the more real data processed in the sample, the more accurate will be the result. That is, by means of numerical simulations it is possible to predict the amount of space relative errors of computation of mathematical expectation as $\dfrac{\Delta E(\nu_i)}{\sigma(\nu_i)}$ a function of the number of data used in the sample for the normal distribution law of values. These dependences at different value of coefficient of trust are displayed in figure 3.

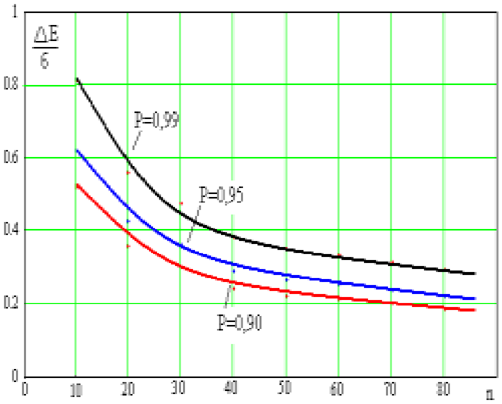

Fig. 3 Nomograms rated on a mean square deviation an interval of mistakes established with confidential probability 0.99, 0,95 and 0.90

Similarly it is possible to receive relative value of an expected interval of an error of calculation of a mean square deviation. Curve communications of an error of a mean square deviation and the amount of studied selection are given in figure 4.

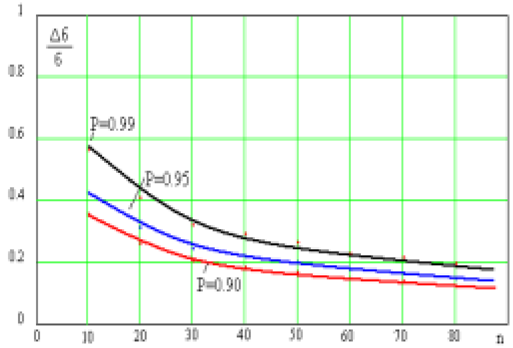

Fig. 4 Relative error of an assessment of a mean square deviation for the intervals corresponding to confidential probability 0.99, 0,95 and 0.90

The third of the younger statistical moments are correlation coefficients. Communication of a relative error of an assessment of value of coefficients of correlation of biometric parameters with the volume of basic data is given in figure 5.

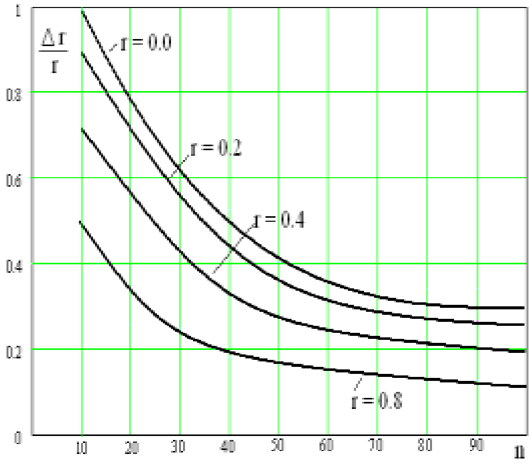

Fig. 5 Relative error of calculation of coefficients of the correlation, caused by final number of examples in test selection at confidential probability 0.99

From the drawings given above it is visible that the error of measurement of the younger statistical moments monotonously decreases, however at the small volumes of test (training) selection of a mistake are considerable and make about 50% of the estimated parameter. So considerable size of mistakes does almost impossible use of classical square forms even rather low order.

For example, if we try to estimate conditionality number for correlation matrixes with identical coefficients of correlation of r = 0.4 (this size is close to average value of modules of coefficients of correlation, see figure 1), we will receive the following values of sequence of numbers of conditionality:

$$\mathrm{cond}\begin{pmatrix} 1 & 0.4 \\ 0.4 & 1 \end{pmatrix} = 2.762, \quad \mathrm{cond}\begin{pmatrix} 1 & 0.4 & 0.4 \\ 0.4 & 1 & 0.4 \\ 0.4 & 0.4 & 1 \end{pmatrix} = 4.819, \quad \mathrm{cond}\begin{pmatrix} 1 & 0.4 & 0.4 & 0.4 \\ 0.4 & 1 & 0.4 & 0.4 \\ 0.4 & 0.4 & 1 & 0.4 \\ 0.4 & 0.4 & 0.4 & 1 \end{pmatrix} = 7.11, \ldots$$

In process of growth of dimension of a correlation matrix the number of its conditionality monotonously grows. It is known that the coefficient of conditionality can be considered as coefficient of strengthening of errors of basic data that is even in attempts to solve two-dimensional problems we have to receive results with a mistake 50%×2.76 ≈ 138%. For this reason in biometrics, biology and medicine it is impossible to use simple and clear linear algebra. Even low-dimensional square forms of linear algebra start working effectively at several hundred examples used for training (the matrix address) or for testing.

If we refuse the usage of classics (2) and we apply big artificial neural networks, for their training in accordance with GOST P 52633.5-2011 [6] it is enough from 20 to 30 examples of an image "Own" and as much examples "Own" and "Others" it is necessary for testing of quality of made decisions [7, 8]. There is no alternative to artificial neural networks at processing biometric, biological, medical, sociological, economic data.

Conclusions and practical recommendations

In case the reader of this article remains within classical one-dimensional statistical processing of biological, biometric, medical and other data on each of observed parameters, the nomogram of figures 3 and 4 will allow estimating an interval of the possible mistakes arising because of final number of examples in observed selection. The material given above is possible to consider as rather simple technique of an assessment of errors of calculation of a population mean and a mean square deviation in a one-dimensional case.

If the reader decides on processing of biological data by more difficult instrument of application of big artificial neural networks, it will have additional technological capabilities. First considerably (on some orders) probabilities of errors of the first and second sort decrease. Secondly there is a possibility of correct increase in the amount of test selection by a butstrap-method or a way of crossing of examples parents and receiving images descendants from them in accordance with GOST P 52633.2-2010 [5]. In the latter case the estimates of errors of measurement of the younger statistical moments given in this work (figures 3 and 4) are not that other as uncertainty present at data. Uncertainty actually considered in this work is no other than a source of casual mutations at reproduction of examples parents. In this regard it is possible to refuse the additional mutations provided by GOST P 52633.2-2010 [5] and components from 1% to 3% of the bio-data. When using small selections of basic data to add from 1% to 3% casual variations there is no sense as the limited volume of selection already provides considerable level of casual components (real uncertainty of the center of ellipsoids and their diameters the right part of figure 2 illustrates).

Another application, discussed in this article uncertainty is the possibility to improve the quality of teaching large artificial neural networks. Knowing the uncertainty range of the mathematical expectation and standard deviation, we can go through the possible close the status of weights training neurons, thus having an improved version of the standard algorithm [7, 8]. Presumably that another area of improvement of learning algorithms will be used when calculating the weights of the neuron of the coefficients of correlation of data. In this case, the uncertainty estimates of the correlation coefficients should be implemented in accordance with the functions listed in figure 5.

References

1. Volchikhin V.I., Ivanov A.I., Funtikov V.A., Nazarov I.G., Yazov Yu.K. (Neural protection of personal biometric data. // Moscow: Radiotekhnika, 2012., 157 s., ISBN 978-5-88070-044-8.
2. Akhmetov B.S., Volchikhin V.I., Kulikov S.V., Malygina E.A. Modeling of the long biometric codes reproducing correlation communications of the output data of the neuronetwork converter. - Moscow: Radiotehnika, Nejrokomp'jutery: razrabotka, primenenie, №3, 2012. S. 40-43.

3. B. Akhmetov, A. Doszhanova, A. Ivanov, T. Kartbayev and A. Malygin "Biometric Technology in Securing the Internet Using Large Neural Network Technology. World Academy of Science, Engineering and Technology. Issue 79, July, 2013, Singapore, p. 129-138, pISSN 2010-376X, eISSN 2010-3778, www.waset.org
4. Boll R.M., Konnel Dzh.H., Pankanti Sh., Ratha N.K., Sen'or Je.U. Rukovodstvo po biometrii. Moskva: Tekhnosfera, 2007. -368 s., ISBN 978-594836-109-3
5. GOST R 52633.2-2010 Protection of information. Technology of information protection. Requirements to formation of the synthetic biometric images intended for testing of means of highly reliable biometric authentication.
6. GOST R 52633.5-2011 Protection of information. Technology of information protection. Automatic training of neural network converters biometrics-access code.
7. GOST R 52633.3-2011 Testing resistance means highly reliable biometric security to attacks selection
8. Akhmetov B.S., Volchikhin V.I., Ivanov A.I., Malygin A.Ju. Testing algorithms of biometric-neural mechanisms of protection of the information. – Almaty: KazNTU im. Satpaeva, 2013. - 152 p. ISBN 978-101-228-586-4, http://portal.kazntu.kz/files/publicate/2014-01-04-11940.pdf