Ensuring the adequacy of neural network models based on the optimization of test volumes

S Kovalevskyy[1], O Kovalevska[1], I Tasić[2], R Cvejić[3] and A Koshevoy[1]

[1]Donbass State Engineering Academy (DSEA), Kramatorsk, Ukraine,
[2]University of Novi Sad, Technical Faculty "Mihajlo Pupin", Zrenjanin, Serbia
[3]University UNION - Nikola Tesla, Faculty of Information Technologies and Engineering, Belgrade, Serbia,

e-mail: kovalevskii@i.ua

Abstract: Ensuring the accuracy and adequacy of mathematical models based on neural networks is an insufficiently studied area of modeling complex multifactor models based on a limited amount of input data. The article presents the results of studies of randomly organized test samples that complement the training set to the total amount of set of precedents provided for the construction of a mathematical neural network model of the object. It is shown that the adequacy of the neural network model is achieved by agreeing on the size of the set and the expected accuracy of the model on a given set of procedures. The formulation of the optimization problem for adequate neural network models is formulated. The results obtained significantly expand the approaches to increasing the reliability of neural network models.

1. Introduction

Mathematical modeling of objects is a necessary and, often, the only possible means for studying and applying for practical purposes knowledge about the features of simulated objects. The central concept of the theory of mathematical modeling is the concept of adequacy [**Error! Reference source not found.**,**Error! Reference source not found.**,**Error! Reference source not found.**]. An argumentated adequacy check ensures that good and practically significant results are obtained. Adequacy of the mathematical model is the correspondence of the results of the computational experiment to the behavior of the real object. This correspondence should be evaluated in terms of the research objectives. Therefore, various approaches to assessing the adequacy of different models are possible. To find out this correspondence for processes characterized by measurable quantities - parameters - it is necessary to compare the parameters of the model and the original under the same conditions. The question of the necessary and sufficient degree of conformity of the object - the original or the adequacy of the model is among the most important in the field of modeling methodology, and the answer to this question characterizes the efficiency of modeling, which reflects the practical utility of the model. Therefore, only the corresponding parameters must be compared to each other and only in the area of the functioning of the object in which it is supposed to be investigated. To verify the adequacy of the mathematical model to the real process, it is necessary to compare the values observed in the course of the experiment with the predictions of the model with certain process parameters [**Error! Reference source not found.**,5,**Error! Reference source not found.**].

In modeling, the researcher should always strive for the most complete and accurate reproduction in the model of the properties and characteristics of the object. The consequence of this is the increase in the complexity of the model, which manifests itself in the number of variables, the number of considered links and influences, and increasing the requirement for the accuracy of the source data.

In mathematical terms, the best combination of completeness-accuracy of the model being created on the one hand and simplicity on the other, almost never succeeds because of the non-formalizability and ambiguity of most of the factors to be taken into account. The deciding factor of efficiency is the mathematical apparatus. At the same time, the best quality or efficiency of any model is achieved as a reasonable compromise between the proximity of the model to the original (adequacy) and the simplicity that provides the possibility and convenience of using the model for its direct purpose, and excessive accuracy of the model in practice is no less harmful than its incompleteness [**Error! Reference source not found.**].

Consistency implies the identical nature of the change of the corresponding parameters, i.e. identical form of the basic properties of functional dependences. With a deeper consideration of this concept, the manifold of possible criteria for checking consistency becomes apparent. Since the comparable parameters in the object's area of operation can take many different values, so that any conclusions about the conformity of their behavior can be made only on the basis of statistical processing of such sets. Therefore, the adequacy is verified using statistical criteria, which can with a certain probability indicate the conformity of the results of the computational experiment to the behavior of the real object in the appropriate conditions.

Mathematical models of mechanical systems and processes are constructed basically as similar deterministic models possessing a common mathematical description with the original. Therefore, for the adequacy of the mathematical model of the behavior of the original - the mechanical system - it is enough to be convinced of the performance of two properties: accuracy and consistency. Accuracy in the tasks of mechanics means that the generalized characteristic of the mismatch of the corresponding parameter of the model and the original should not be greater than the predetermined value of the acceptable error. As such, the meanings of the mismatch, the mean value of the mismatch, or the statistical estimate may be the highest modulus.

When identifying difficult formalizable complex systems, when artificial neural networks are used to construct mathematical models, the use of statistical criteria is difficult and sometimes impossible [**Error! Reference source not found.**,**Error! Reference source not found.**].

2. Basic part

An approach based on a neural network basis is a mathematical structure with a variable structure dependent on input data. The structure of the elements of such structures is rebuilt using custom threshold elements whose level of excitation depends on the input data. The verification of the adequacy of such models is based on satisfactory testing on a test set that is not included in the training set, and the volume of the source data arrays is directly related to the quality of the models formed on the basis of these data. The purpose of the research: to establish the relationship between the teaching and test sets in the creation of adequate models of objects of different nature.

In general terms, there is a vector set *X*, which may be represented n by parameters of table 1.

In accordance with the goal set to explore the features of the choice of the trainee *Y* and test *Z* sets, when X϶Y,Z a table of precedents was formed (table 2).

The 32 tuples of the set *X* (table 2), test samples were created, representing 9 random combinations with 9 levels of volumes of test sample arrays C_Y^Z :

$$C_{32}^{3};\ C_{32}^{6};\ C_{32}^{9};\ C_{32}^{12};\ C_{32}^{15};\ C_{32}^{18};\ C_{32}^{21};\ C_{32}^{24};\ C_{32}^{27}\ . \qquad (1)$$

These are tables of combinations of subsets *Y* and *Z*.

Table 1 - Vector set of source data for modeling

№	X_1	X_2	X_3	X_4	X_5	X_6	…	X_n

1	x_{11}	x_{21}	x_{31}	x_{41}	x_{51}	x_{61}	…	x_{n1}
2	x_{12}	x_{22}	x_{32}	x_{42}	x_{52}	x_{62}	…	x_{n2}
3	x_{13}	x_{23}	x_{33}	x_{43}	x_{53}	x_{63}	…	x_{n3}
…	…	…	…	…	…	…	…	…
m	x_{1m}	x_{2m}	x_{3m}	x_{4m}	x_{5m}	x_{6m}	…	x_{nm}

Table 2 - A set of precedents used for neural network simulation

№	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	Y
1	1	0	1	1	0	0	1	0	1	0	0	0	2
2	0	0	0	0	1	0	0	1	1	0	0	0	1
3	1	1	1	0	0	0	1	1	1	0	1	0	1
4	1	1	0	0	1	0	1	0	0	0	1	0	1
5	1	1	0	1	0	1	0	0	0	1	0	0	2
6	1	0	0	1	0	0	1	1	0	0	0	0	1
7	1	0	0	1	0	0	1	0	1	0	1	0	2
8	0	0	0	0	1	0	0	0	0	0	0	0	1
9	0	0	1	0	1	0	0	1	1	0	0	1	2
10	0	0	0	1	0	1	0	1	1	0	1	0	2
11	0	1	0	0	1	0	1	0	0	0	0	1	1
12	1	1	0	0	1	0	0	0	0	0	1	0	1
13	1	1	0	0	0	0	0	1	0	0	0	1	1
14	1	1	1	1	1	0	1	0	0	0	0	0	2
15	0	0	0	0	1	0	0	1	0	0	0	0	1
16	1	0	0	1	0	0	0	1	1	0	0	0	2
17	0	1	1	0	1	0	1	1	0	1	0	0	1
18	1	1	0	0	0	0	1	0	0	0	0	0	1
19	1	1	0	0	1	1	0	0	1	0	0	1	2
20	0	1	0	0	1	1	1	1	0	0	1	0	1
21	1	1	0	0	1	1	1	1	0	0	1	0	1
22	1	1	0	0	1	0	1	1	0	0	1	0	1
23	1	1	1	0	1	0	0	1	0	0	0	0	1
24	1	0	0	1	0	0	1	0	0	1	0	1	2
25	0	1	0	0	1	0	0	0	0	0	1	0	1
26	1	1	0	0	0	1	0	0	0	0	0	1	2
27	0	0	0	0	1	0	1	0	0	0	0	0	1
28	1	1	1	1	0	0	1	1	1	0	0	0	2
29	0	0	0	0	1	0	1	1	1	0	0	0	1
30	1	1	0	1	1	0	0	0	0	1	0	0	2
31	0	0	1	1	1	1	0	0	0	1	0	1	2
32	0	1	1	1	1	1	0	1	1	1	0	0	2

3. Methodology of experimental studies

The research methodology involves conducting neuro-net model accuracy-adequacy estimates based on the variation of the random selection of training and test subsets formed from the same data set available for the creation of a non-grid model. The method of comparative analysis included the following steps:

1. Build neural network models based on neural network modeling and data analysis Neuro Pro 0.25 [**Error! Reference source not found.**] based on data from training subsets *Y*.

2.Creation of the number of inputs of the neural network model x1..x12, leaving only significant values for the training subsets to be obtained for the given accuracy.

3. Testing of neural network models obtained as a result of the implementation of item 2 on test sets *Z* at the same time assessing the accuracy of testing.

4. Building the dependencies of the given training accuracy and the obtained accuracy of testing on the sets of samples of subsets of *Y* and *Z* arrays (1).

5. Analysis of the obtained results and development of recommendations on the method of assessing the adequacy of neural network models.

Thus, 81 neural network models with a given simulation error of ±1.0% were created and analyzed. The results of the studies are presented in table 3.

Figure 1 shows a graphical interpretation of the error *S* of the reproduction of the test set *Z*.

Table 3 - Error of testing 81 samples of neural network models

Accuracy testing	Error of testing with share p-test set from the size of the data set,%									Total error
	10	20	30	40	50	60	70	80	90	S
10,00%	0,335	0,031	0,669	0,002	0,337	0,008	0,004	0,335	0,008	1,727
20,00%	0,282	0,006	0,337	0,195	0,330	0,170	0,171	0,174	0,339	2,003
30,00%	0,118	0,198	0,008	0,340	0,401	0,117	0,113	0,170	0,005	1,472
40,00%	0,202	0,124	0,173	0,170	0,179	0,173	0,090	0,171	0,257	1,539
50,00%	0,401	0,536	0,138	0,378	0,138	0,213	0,177	0,140	0,205	2,325
60,00%	0,173	0,447	0,113	0,501	0,214	0,225	0,348	0,560	0,325	2,906
70,00%	0,146	0,656	0,287	0,435	0,170	0,226	0,351	0,101	0,312	2,684
80,00%	0,500	0,737	0,253	0,169	0,337	0,336	0,379	0,424	0,382	3,517
90,00%	0,408	0,371	0,151	0,334	0,446	0,335	0,721	0,150	0,482	3,398

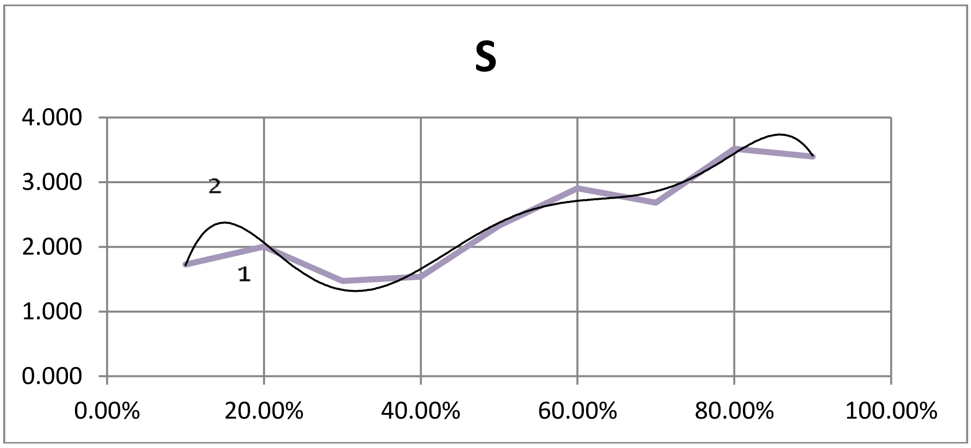

Figure 1 - Total Error 1 and its approximation by the polynomial of 2 test variants with a fraction of the test set from 10% to 90%.

Analysis of simulation and testing errors of models presented by the trend curve approximated by the polynomial dependence of the form:

$$\mathbf{S} = -2219{,}9p^6 + 6889{,}8p^5 - 8388{,}4p^4 + 5041{,}9p^3 - 1535{,}2p^2 + 216{,}97p - 8{,}898 \quad (2)$$

allowed to determine the minimum of error.

The verbal description [**Error! Reference source not found.**] of the optimized model is represented by the following description (3):

Database fields (initial symptoms):

X4
X5
X8
X9
X12

Database fields (end syndromes):

y

Preprocessing the input fields of the database for supplying the network:

X4=(X4-0,5)/0,5
X5=(X5-0,5)/0,5
X8=(X8-0,5)/0,5
X9=(X9-0,5)/0,5
X12=(X12-0,5)/0,5

Functional Converters:

Sigmoid 1(A)=A/(0,1+|A|)

Level 1 syndromes:

Syndrome1_1=Sigmoid 1(0,5175867*X4-0,2253383*X5-0,7707164*X8+0,167638*X9-0,1165028*X12)
Syndrome1_2=Sigmoid 1(0,1908794*X8-0,09852738*X9+0,1385741*X12-0,1664897)
Syndrome1_3=Sigmoid 1(0,3572189*X4+0,3209354*X12)
Syndrome1_4=Sigmoid 1(0,6674076*X4-0,115299*X5-0,8144038*X8+X9+0,1843456*X12+0,7063498)

End-syndromes:

y=0,6864882*Syndrome1_1+Syndrome1_2-0,4366252*Syndrome1_3+Syndrome1_4+0,3727925

Post-treatment of end-syndromes:

Y=((y*1)+3)/2)

The minimum of the total error, which was estimated with respect to randomly formed subsets Y and Z came upon setting the test set as 1/3 of the randomly selected portion of the common data array. This recommendation, of course, differs from the recommendations of 80% / 20% for the fraction of subsets Y and Z [**Error! Reference source not found.**] and should be taken into account when constructing neural network models.

4. Conclusions

The studies, the results of which are presented in the article, showed that for neural network models, adequacy and accuracy are achieved under the condition of formation of the training and test sample, proceeding from the achievement of satisfactory accuracy at the maximum volume of the test sample. Therefore, the minimum of the total error at 10% of the fraction of the test set indicates a high probability of recognizing the test set. The minimum of the total error of the test set must correspond to condition (3):

$$\mathbf{S} \rightarrow \min; \mathbf{p} \rightarrow \max \qquad (3)$$

Applying this rule can ensure the accuracy and adequacy of the neural network model, based on the optimal amount of training and test sets. Neglecting this rule will result in the loss of the adequacy of the model with the ultimate achievable accuracy. It should be noted that the accuracy of the neural network model will be limited by the volume of the set of precedents.

References

[1] U.B. Filik, M. Kurban. A new approach for the short-term load forecasting with autoregressive and artificial neural network models. *International Journal of Computational Intelligence Research*. 3 (2007) 66-71.

[2] O.S. Kovalevska, S.V. Kovalevskyy. Application of acoustic analysis in control systems of robotic machine tools. *Naukovij zhurnal «Radioelektronika, informatika, upravlinnya» - «Radio Electronics, Computer Science, Control»*. 2 (45) (2018) 51-59.

[3] S. Kovalevskyy, O. Kovalevska, R. Turmanidze. Acoustic diagnostics of lever mechanisms with subsequent processing of data on neural networks. *Lecture Notes in Networks and Systems (LNNS)*. 42 (2019) 202-210.

[4] I.O. Konovalova, Yu.A. Berkovich, A.N. Erohin. Optimization of the LED lighting system vitamins space greenhouse. *Aviakosmicheskaja i jekologicheskaja medicina*. 50 (3) (2016) 17-22.

[5] A.E. Kononyuk. *Obobshhennaya teoriya modelirovaniya*. Nachala. K.1.Ch.1 :"Osvita Ukraïni", 2012. - 602 pp.

[6] S. Kovalevskyy, O. Kovalevska. Resource optimization with systemic design of robotized technological equipment. In: *Proceedings of the World Convention on Robots, Autonomous Vehicles and Deep Learning, Singapore*, 10-11 September 2018. 2018, pp. 50. doi: 10.4172/2168-9695-C3-0216th.

[7] V.A. Kaladze. *Mathematical models of casual processes with stationary increments and the non-uniform information dynamic processing*. Monograph. Lorman (MS - USA): Science Book Publishing House, 2012. - 136 pp.

[8] E.A. Ganceva, V.A. Kaladze, A.M. Polyakov. Intellektual'ny'j kriterij kachestva matematicheskix modelej slozhny'x sistem: ideologiya, perspektivy' razrabotki. *Vestnik Voronezhskogo gosudarstvennogo texnicheskogo universiteta*. 9 (5.1) (2013) 52-56.

[9] V.G. Caregorodcev. Izvlechenie yavny'x znanij iz tablic danny'x pri pomoshhi obuchaemy'x i uproshhaemy'x iskusstvenny'x nejronny'x setej. In: *Novy'e informacionny'e texnologii : materialy' vtorogo nauch.-prakt. seminara*. Moscow: Mosk. gos. in-t e'lektroniki i matematiki (MGIE'M), 1999. pp.. 40-50.

[10] A.N. Gorban', D.A. Rossiev. *Nejronny'e seti na personal'nom komp'yutere*. Kiev (Ukraine): Nauka, 1996. - 276 pp.