UDC 004.942
DOI: 10.32626/2308-5916.2023-24.45-55

Valentyn Krykun, Graduate Student

Odessa Polytechnic National University, Odessa

IMPROVING THE ACCURACY OF THE NEURAL NETWORK MODELS INTERPRETATION OF NONLINEAR DYNAMIC OBJECTS

The paper is devoted to the problem of neural network interpretation in the tasks of modeling nonlinear dynamic objects. The purpose of the work is to improve the accuracy of the neural network models interpretation of nonlinear dynamic objects and to determine the scope of their effective application. This goal is achieved by applying analytical models in the form of integral-power series based on multidimensional weight functions. The scientific novelty of the work lies in the use of nonlinear dynamic models in the form of integral-power series based on multidimensional weight functions instead of linear surrogate models. It allows to improve modeling accuracy. The practical usefulness of the work is determination of the effective application area of analytical interpretive models. The practical significance of the obtained results lies in the application of the proposed models for the interpretation of neural network models of nonlinear dynamic objects, which allows to increase the accuracy of interpretation models compared to linear surrogate models.

Keywords: *interpretation of machine learning models, nonlinear dynamic models, time delay neural networks*.

1. Introduction. The development of science and technology makes it possible to ensure a qualitative increase in the characteristics of modern devices and processes in various fields of activity. On the other hand, this continuous process leads to the constant complexity of control objects, and toughening requirements for their functioning [1].

For successful interaction with such objects (solving problems of control, management, diagnostics), it is first of all necessary to provide their adequate mathematical support and effective modeling tools. This is achieved primarily by increasing the complexity of models [2].

Thus, in the last decade, there has been a significant leap in the computing power of computer technology, combined with the development of information processing algorithms and access to large amounts of data. This has resulted in significant progress in machine learning, which has led to an increase in modeling accuracy and, as a result, the widespread use of machine learning models in various fields of activity. At the same time, the accuracy of machine learning models with their ability to automatically detect, study and obtain useful knowledge from large amounts of data is achieved by increasing the complexity of the model itself. As a result, the ability to explain the principles of operation of such a model decreases and, as a result, the results obtained become quite difficult for humans to understand, i.e., the interpretability of the model is impaired.

Interpretability is an important property of machine learning models. It facilitates the processes of controlling and diagnosing object by explaining why a particular solution was obtained, which helps to improve the model.

In a number of industries where there are increased requirements for model adequacy, namely, where the model can significantly affect people's lives, such as medicine, finance, transportation, cybersecurity, interpretability is not only a desirable property of the model, but also an inherent requirement for machine learning models, enshrined in law (for example, the European GDPR regulation [3], which requires the right to explain the decision made by the model).

2. The research purpose and problem formulation. Not all machine learning models require interpretation. There is a trade-off between model interpretability and model complexity. For example, regression models, shallow decision trees, and k-nearest neighbors models (in the space of interpretable features) are simple models and are easily perceived by humans, i.e., interpretable. On the contrary, neural networks and gradient boosting are models whose working principle is difficult to understand, so these models require additional interpretation [4].

Of greater scientific and practical interest are the tasks of modeling complex objects with unknown laws of functioning and unknown structure, when the use of simple interpretable models does not lead to a satisfactory result. Such objects are usually considered as a «black box» [2, 3].

As examples of black box objects, we can consider nonlinear dynamic objects with unknown laws of functioning and unknown structure. Due to the nonlinear dynamic characteristics, object can function in more complex modes that cannot be realized using linear characteristics [2]. Such objects are characterized by some a priori uncertainty: lack of data about the objects, the presence of interference and disturbances in the external environment. Therefore, traditional deterministic methods are not suitable for modeling such objects.

When modeling such objects, the neural network approach is becoming more widespread. Neural networks have gained popularity due to the fact that the process of building a model requires only measuring the data at the input and output of the «black box» and does not require any assumptions about the structure of the object and the internal laws of its functioning. Therefore, the use of neural networks to describe non-linear dynamic objects, in particular, those with continuous characteristics, has recently expanded significantly.

However, due to the high nonlinearity and complex interactions of a large number of model parameters, neural networks do not explicitly reflect the structure and internal laws of the object's functioning. Therefore, neural networks are perceived as «black box» models. As a result, a serious disadvantage of neural network models is that the predictions made by such a complex model cannot be traced back to the input data and understand why the output data is transformed in a certain way.

Thus, the complex interactions of a large number of parameters in neural networks are not easy to trace, while disentangling them can provide insight into the processes reflected in the object and the parameters on which the model's decisions are based. As a result, models in the form of neural networks do not provide a clear analytical expression of the relationship between the input and output of an object. At the same time, it is convenient to use an analytical model to analyze the properties of black box objects, which allows to draw clear and unambiguous conclusions about the functioning of the system.

As a result, when modeling black-box objects, one has to deal with models whose operating principles are not obvious, and whose features often do not have a physical meaning, which complicates the interpretation of models by humans. As a result, the widespread use of neural network models is significantly constrained in such critical areas as medicine, finance, and transportation, where there are increased requirements for the security of modeling results and model credibility.

On the other hand, the situation that has developed in recent years has stimulated interest in research on the interpretation of black boxes based on neural networks in order to increase confidence in these models, analyze the structure and laws of functioning of the objects under study [2, 3]. Therefore, the development of the interpretation of neural network models of nonlinear dynamic objects remains an urgent task.

The purpose of the work is to improve the accuracy of the neural network models interpretation of nonlinear dynamic objects and to determine the scope of their effective application.

3. Literature overview. Today, the greatest efforts of scientists and practicing engineers are concentrated in the area of machine learning model interpretation. World-renowned IT companies also see great potential in this area and create their own tools for interpreting machine learning models: aix360.readthedocs.io (IBM), aws.amazon.com/sagemaker (Amazon), captum.ai (Facebook), explainable.ai (Google), interpret.ml (Microsoft), etc.

The analysis of scientific achievements and the above services allows us to distinguish the following approaches to the interpretation of machine learning models:

- visualization (2D, 3D graphs, graphs, etc.);

- textual explanation (models in the form of antecedents: «factor A and factor B led to prediction C»);
- numerical estimates (importance of features, coefficients, weights);
- analytical expressions (explicit dependence of an object's output on its input).

When interpreting neural networks, approaches based on visualization and evaluation of the importance of features are most often used [3, 4].

Visualization. Methods of interpreting neural networks based on visualization of decision-making processes for processing non-numerical data: images and video, sound and speech, text [3] are widely known.

Advantages: visual display of the initial data in the image space, assessment of the quality of the machine learning process of the model.

Disadvantages: lack of numerical estimates of the relationship between features and their importance, explicit dependence of the object's output on its input.

Numerical estimates. In the interpretation of neural networks, methods based on the assessment of the importance of features are often used to explain individual model predictions. A popular method for assessing the importance of features is the SHAP (SHapley Additive exPlanations) method [5].

Advantages: assessment of the importance of features, identification of features that affect the model performance.

Disadvantages: lack of assessment of the functional dependence of the result on the identified features, the explicit dependence of the output of the object on its input.

Analytical expressions are used to interpret models much less frequently, although they have a number of important advantages over other approaches. Interpretation of models in the form of analytical expressions allows to ensure mathematical reliability: to transparently show the absence of hidden behavior or logic that affects the behavior of the model [3, 4].

The mathematical support for constructing interpretable analytical input/output expressions for machine learning models is not fully developed and is usually reduced to approximating the model in the local domain with simpler computationally simple surrogate models, in particular, linear models [4]. Popular analytical methods for interpreting machine learning models are LIME (Local Interpretable Model-agnostic Explanations) and linear regression [3], which build surrogate models by locally approximating the original model to a linear one.

Among the machine learning models, it is important to distinguish neural network models. The relevance of the task of interpreting neural networks using analytical expressions is increasing due to the fact that neural networks capable of carrying reliable information about the structure and functions of the control object are increasingly used to model complex objects and processes of the world around us (technical and biological control objects, production, control and automation facilities). Nevertheless, neural network models of such control systems have not been sufficiently studied, and methods for interpreting these models are much less represented in the literature and are usually reduced to linearization [4] or polynomial approximation [3]. At the same time, the interpreting models usually take the form of linear dynamic or nonlinear static dependencies and do not reflect all the properties of the object.

This problem can be solved if nonlinear dynamic models, such as integral-power series based on multidimensional weight functions, are used as interpretation models [6]. The main advantages of these models are the simultaneous consideration of nonlinear and dynamic properties of the object, which ensures an increase in the accuracy of the interpretation of neural network models of nonlinear dynamic objects.

As a result of an analytical review of the current state of the problem of interpreting neural network models, the paper proposes an approach to building interpretive models based on an analytical expression in the form of integral-power series based on multidimensional weight functions. The use of this approach allows to simultaneously increase the accuracy and reduce the computational burden of interpretation of complex research objects.

In this work, the method of interpreting models of complex objects with nonlinear and dynamic properties of the "black box" type in the form of neural networks was further developed by using the analytical expression of the input/output relationship of the model in the form of integral-power series based on multidimensional weight functions.

4. Main part.

4.1. Simulation model of the test object. Using the example of a test object, we investigated the effectiveness of interpreting non-network models. The simulation model of the test object with a first-order dynamic block and a nonlinear feedback block [7] is shown in Fig. 1.

The polynomial function $f(y) = y^2$ is considered as a nonlinear function $f(y)$ in the feedback block. The simulation model is studied using test signals with different amplitudes: pulse, step, linear, and harmonic.

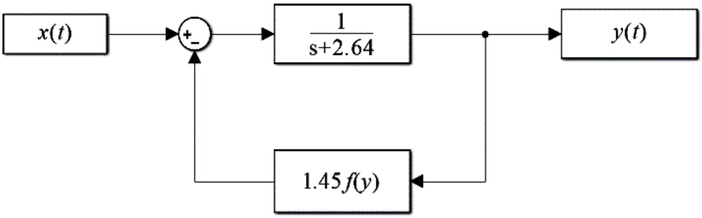

Fig. 1. *Simulation model of a test nonlinear dynamic object*

In Fig. 2 shows the transient characteristics $y_{03}(t)$, $y_{06}(t)$, $y_{09}(t)$ of the test object when the input is a step signal $x(t) = a\Theta(t)$, $a = 0.3, 0.6, 0.9$. The figure demonstrates test object`s nonlinear and dynamic properties.

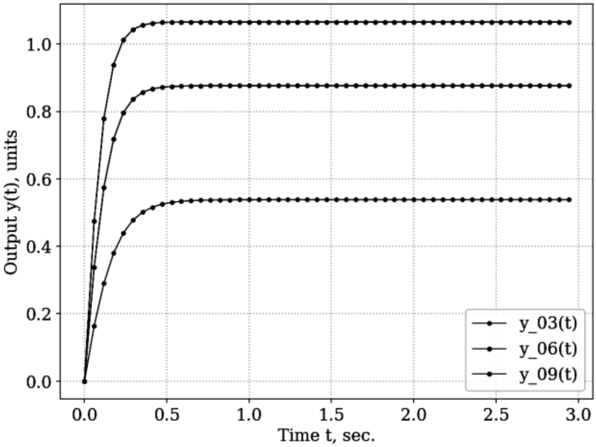

Fig. 2. *Transient characteristics $y_{03}(t)$, $y_{06}(t)$, $y_{09}(t)$ of the test object under the action of a step signal at the input*

4.2. Object model based on a neural network with a time delay. There are several neural network structures for modeling nonlinear dynamics: Dynamic Neuro-SM, Wiener-type DNNs, and time-delay neural networks (TDNNs) [8].

Among these types of neural network models, TDNNs are the most general structure consisting of several layers with direct signal propagation [6-8]. Such models are able to learn from the input-output data of nonlinear dynamic objects and have excellent convergence properties [7], which are advantages over the aforementioned Dynamic Neuro-SM and Wiener-type DNN methods. Therefore, TDNN models are an effective tool for modeling nonlinear dynamic objects with continuous characteristics.

In this paper, a time-delay neural network is used to build an object`s model. Most often, in practice, a three-layer TDNN structure is used with layers: input, hidden, and output.

The signal $y(t_n)$ of the TDNN model is described by the expression:

$$y(t_n) = b_0 + s_0 \sum_{i=1}^{K} w_i S_i \left(b_i + \sum_{j=1}^{M} w_{i,j} x(t_{n-j}) \right), \tag{1}$$

where M is the memory length of the object model, K is the number of neurons of the received layer with a nonlinear activation function, b_0, b_i is the bias of the neurons of the output and hidden layers, respectively; S_0, S_i are the activation functions of the neurons of the output and input layers, respectively; w_i, $w_{i,j}$ are the weighting coefficients of the neurons of the output and hidden layers, respectively.

A TDNN model can be trained to behave dynamically with the incorporation of nonlinear characteristics [7] on the input-output data.

The training data set is formed from the results of the input-output experiment - a set of vectors $\{x(t), y(t)\}$ for each type of input signal.

Based on the results of training the neural network model, the following values of M, K, w_i, $w_{i,j}$ and b_0, b_i, are found, which ensure sufficient modeling accuracy. However, the values of w_i, $w_{i,j}$ and b_0, b_i have no physical meaning and model (1) is difficult to interpret.

Using the interpretation approaches discussed above, in the next section, we build interpretive models for (1).

4.3. Building interpretive models.

4.3.1. Graphical interpretation. This interpretation approach allows visualizing the neural network training process and training results in the following forms:

- scalar functions in the context of time $c(t) = f(\hat{y}(t))$, where $\hat{y}(t)$ is the estimate of the output signal of the object. The loss function or cost function and the accuracy of the neural network are most often used as $c(t)$;
- images formed on data in 2D space $c(a, b) = f(a, b)$. As c, we can use the network weights $w_{i,j}$;
- model graph $c(b_i, w_{i,j}) = \{b_i, w_{i,j}\}$;
- histogram of changes in the distribution of data in layers over time $c(a) = \{a_i\}$, $i = 1, n$, data dimension.

To demonstrate the graphical interpretation of the neural network model of the test object, we use histograms of the time distribution of data in the input (Fig. 3) and hidden (Fig. 4) layers.

Graphical interpretive models can be used to establish qualitative indicators of the significance and relatedness of the properties of a test object. However, these models do not provide information on how certain properties affect the model's predictions.

4.3.2. Numerical estimates of the model. This approach to interpretation allows us to evaluate the process of neural network training and the results of training in the following forms:

- calculation and visualization of the estimates of the significance and connectivity of the features;
- textual explanation (models in the form of anti-cases: «if the feature x_i changes/increases, the probability of a decrease/increase in the prediction $y(x)$ increases»).

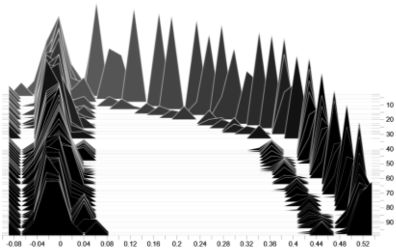

Fig. 3. *Histogram of data distribution in time in the input layer*

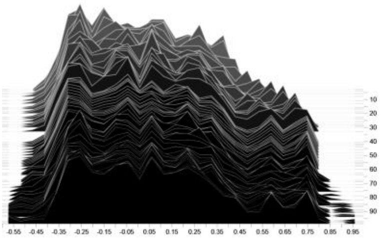

Fig. 4. *Histogram of data distribution over time in a hidden layer*

To analyze and interpret the neural network model of the test object, the SHAP method builds a graph of the importance of the features used in the model (Fig. 5).

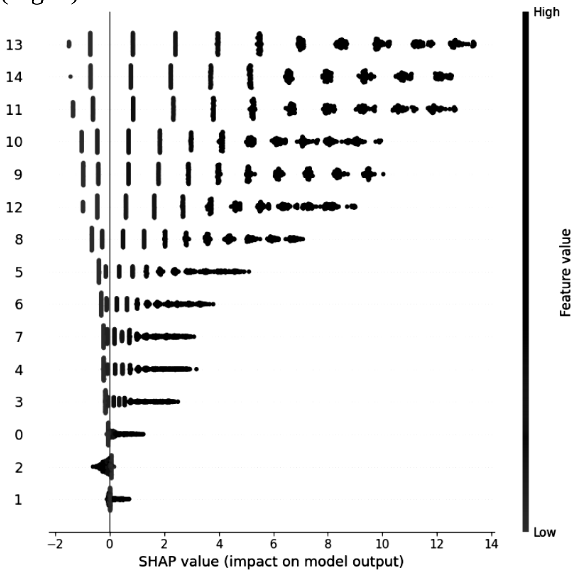

Fig. 5. *Graph of importance of test model features*

Thus, important conclusions can be drawn from this graph and their adequacy can be checked:

- almost all the features are interrelated;
- the features x_i with lower ordinal numbers (located closer to the beginning of the vector $x(\underline{t})$) have greater significance (influence on the output value $y(\underline{t})$).

4.3.3. Analytical model. For a wide class of nonlinear dynamic objects with continuous characteristics, the relationship between the input

$x(t)$ and output $y(t)$ signals can be written in the form of an integral-power series based on multidimensional weight functions [6, 7]. Thus, for an object with one input and one output in the time domain, the model takes the following form:

$$y(t)=\sum_{n=0}^{\infty}\int_0^t\ldots\int_0^t w_n(t,\tau_1,\ldots,\tau_n)\prod_{i=1}^{n}x(\tau_i)d\tau_i \ , \qquad (2)$$

where $x(t)$ and $y(t)$ are the input and output signals of the object; $w_n(\tau_1, \ldots, \tau_n)$ are multidimensional weight functions of the n^{th} order (n =1,2, 3, ...); w_0 is the free term of the series; t is the current time.

To interpret the neural network model of a test object in the form of an analytical expression of an integral-power series, expressions that establish an analytical relationship between these models are used [7]. These expressions are used to determine the multidimensional weight functions of the first $w_1(\tau_1)$ and second $w_2(\tau_1, \tau_2)$ orders.

Fig. 6 shows the multidimensional weighting function of the first order $w_1(\tau)$ and the diagonal section of the multidimensional weighting function of the second order $w_2(\tau, \tau)$.

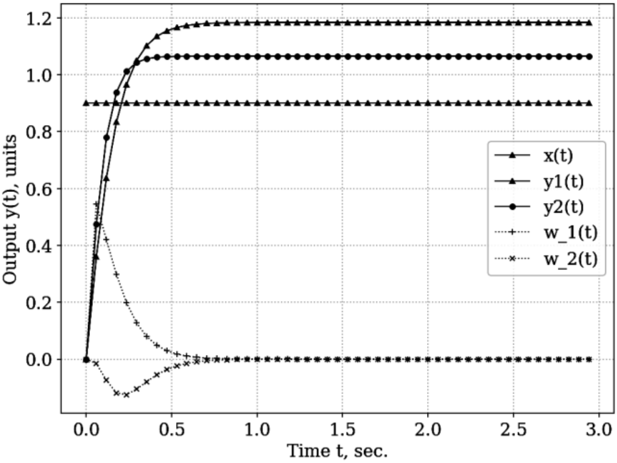

Fig. 6. *Multivariate weighting function of the first-order test object $w_1(\tau)$, diagonal intersection of the second-order multivariate weighting function $w_2(\tau, \tau)$ and interpretive models of the first $y_1(t)$ and second $y_2(t)$ order*

Based on the determined multidimensional weighting functions, the neural network interpretation model is built in an analytical form according to expression (2). The resulting interpretation model transparently demonstrates the structure of the test object and the functional relationship of the features that affect its behavior. To determine the accuracy of the built model, Fig. 6 also shows a linear interpretation model built by the

LIME method. The interpretation model built by expression (2) demonstrates an accuracy of 10-12% higher than the linear model when using test input signals with amplitudes $a \in [0.6, 0.9]$.

5. Conclusion. The paper deals with applied aspects of improving the accuracy of interpreting neural network models of nonlinear dynamic objects. As a result of an analytical review of approaches to the interpretation of neural networks: visualization, numerical estimates of features and analytical expressions, the areas of effective application of analytical interpretation models are identified. Thus, approaches to the interpretation of machine learning models based on visualization and numerical estimates of features do not allow the construction of an analytical expression of the input/output relationship of the control object. Therefore, when modeling complex «black box» objects with nonlinear and dynamic properties, it is advisable to use interpretive analytical models in the form of integral-power series based on multidimensional weight functions to ensure the absence of hidden behavior or logic that affects the behavior of the model.

On the example of a test nonlinear dynamic object, a model in the form of a neural network with time delays is constructed. For the obtained model, an interpretive model in the form of integral-power series based on multidimensional weight functions is constructed, which allows simultaneously increasing the modeling accuracy by 10-12% compared to linear interpretive models and reducing the computational burden of interpreting complex research objects compared to neural networks with time delays.

References:

1. Agresti A. Foundations of linear and generalized linear models. *Wiley series in probability and statistics.* 2017.
2. Schoukens J., Ljung L. Nonlinear System Identification: A User-Oriented Road Map. *IEEE Control Systems Magazine.* 2019. Vol. 39. № 6. P. 28-99.
3. Md. Rezaul Karim, Md Shajalal, A. Grass. Interpreting Black-box Machine Learning Models for High Dimensional Datasets. *arXiv preprint.* 2022. arxiv.org/abs/2208.13405.
4. Cinar A. Overview of existing approaches for the interpretation of machine learning models. *Hochschule Esslingen.* 2019. P. 1-11.
5. Lundberg S. A Unified Approach to Interpreting Model Predictions. *Conference on Neural Information Processing Systems (NIPS 2017).* Long Beach, CA, USA. P. 1-10.
6. Stegmayer G., Pirola M., Orengo G., Chiotti O.. Towards a Volterra series representation from a neural network model. *WSEAS Transactions on Circuits and Systems.* 2004. Archive 1. P. 55-61.
7. Fomin O., Polozhaenko S., Krykun V., Orlov A., Lys D. Interpretation of Dynamic Models Based on Neural Networks in the Form of Integral-Power Series / *Smart Technologies in Urban Engineering. STUE 2022. Lecture Notes in Networks and Systems.* 2023. Vol 536. P. 258-265.

8. Liu W., Su Y., Zhu L. Nonlinear Device Modeling Based on Dynamic Neural Networks: A Review of Methods. *IEEE 4th International Conference on Electronic Information and Communication Technology (ICEICT).* Xi'an, China. 2021. P. 662-665.

ПІДВИЩЕННЯ ТОЧНОСТІ ІНТЕРПРЕТАЦІЇ НЕЙРОМЕРЕЖЕВИХ МОДЕЛЕЙ НЕЛІНІЙНИХ ДИНАМІЧНИХ ОБ'ЄКТІВ

Робота присвячена проблемі інтерпретації нейронних мереж в задачах моделювання нелінійних динамічних об'єктів. Метою роботи є підвищення точності інтерпретації нейромережевих моделей нелінійних динамічних об’єктів та визначення сфери їх ефективного застосування.

В результаті аналітичного огляду підходів до інтерпретації нейронних мереж: візуалізація, числові оцінки ознак та аналітичні вирази, – визначено сфери ефективного застосування аналітичних інтерпретаційних моделей. Так, підходи до інтерпретації моделей машинного навчання на основі візуалізації та числові оцінки ознак не дозволяють збудувати аналітичного виразу залежності «вхід/вихід» об'єкту контролю. Тому, при моделюванні складних об'єктів типу «чорна скриня» з нелінійними і динамічними властивостями для забезпечення підвищених вимог до безпеки результатів моделювання (для переконаності у відсутності прихованої поведінки чи логіки, які впливають на поведінку моделі) доцільно використовувати інтерпретаційні аналітичні моделі у вигляді інтегро-ступеневих рядів на основі багатовимірних вагових функцій.

На прикладі тестового нелінійного динамічного об'єкту збудовано модель у вигляді нейронної мережі з часовими затримками. Для отриманої нейромережевої моделі збудовано інтерпретаційну модель у вигляді інтегро-ступеневих рядів на основі багатовимірних вагових функцій, яка дозволяє забезпечити одночасно підвищення точності моделювання на 10-12% у порівнянні з лінійними інтерпретаційними моделями та зменшення обчислювального навантаження інтерпретації складних об'єктів дослідження у порівнянні з нейронними мережами з часовими затримками.

Ключові слова: *інтерпретація моделей машинного навчання, нелінійні динамічні моделі, нейронні мережі з часовими затримками.*

Отримано: 26.10.2023