

UDC 681.327.12(045)

DOI:10.18372/1990-5548.64.14852

¹O. I. Chumachenko,²A. T. Kot,³A. E. Mandrenko

ALGORITHM OF HYBRID GMDH-NETWORK CONSTRUCTION FOR TIME SERIES FORECAST

^{1,2,3}Technical Cybernetic Department, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute,” Kyiv, Ukraine

E-mails: ¹chumachecko@tk.kpi.ua ORCID 0000-0003-3006-7460, ²anatoly/kot@gmail.com, ³anya.mand@gmail.com

Abstract—It is considered the problem of structural-parametric synthesis of a hybrid neural networks based on the use of Group Method of Data Handling neural network. Hybridization is achieved through the use of various neurons: classical, nonlinear Adaline, R-neuron, W-neuron, Wavelet-neuron. The problem of structural-parametric synthesis of hybrid neural network consists in the optimal choice of the number of layers, the number of neurons in the layers, the order of alternation of layers with different neurons. As an example it is considered the forecast problem solution with help of hybrid neural networks based on the data of the COVID-19 pandemic, collected by Johns Hopkins University. A MAPE criterion was used for quality assessment.

IndexTerms—Hybrid neural network; structural-parametric synthesis; forecast of time series.

I. INTRODUCTION

The need to forecast the time series arises in cases where it is necessary to make decisions based on accumulated data on the system behavior: it can be the control which has inertia, analysis of financial markets for effective investment of money, of a complex apparatus with inertia, making decision under managing an enterprise or business that depends, for example, from the expected profit, etc.

One such case it can be seen now. Due to the COVID-19 pandemic, the power of Ukraine and other countries are forced to decide about reduction or introduction the quarantine measures of epidemic containment. The decision in these cases is made taking into account the dynamics of the spread of the virus. Correct forecast of morbidity, mortality and economic indicators that depend on quarantine restrictions can help make the best decision. The lives and material wealth of millions of people depend on the quality of forecasts, so it is difficult to overestimate the relevance of this problem. One of the ways to solve the problem of forecast is the use of artificial neural networks, especially hybrid neural networks (HNN) [1].

A hybrid neural network is a network consisting of several neurons or neural networks that have different properties or solve different problems. The combination of such topologies in a certain order can lead not only to a combination of their advantages, but also to the effect of synergy, i.e. obtaining new useful properties. Structural-parametric synthesis of HNN in this case includes not only the selection of the number of network layers and the number of

neurons, but also the sequence of network architectures or combinations of neurons of different types within one architecture [3].

II. PROBLEM STATEMENT

To solve the structural-parametric synthesis of HNN, it is first necessary to determine which artificial networks and neurons can be used to construct the hybrid network.

The paper proposes to use one of the known topologies of neural networks and to saturate it with neurons of different types [2], [4] – [6].

The problem of structural-parametric synthesis of HNN is posed, which consists in optimal choice of basic neural network topology, number of layers, number of neurons in the layers, order of layers alternation with neurons of different types [3].

III. PROBLEM SOLUTION

It is proposed to choose a multilayer evolutionary Group Method of Data Handling (GMDH) neural network for its saturation with neurons of different types during training [4], [5]. The selection of neurons for each layer will be made automatically by an algorithm based on the criterion of minimum error. It's developed an algorithm for constructing an input sample and an updated algorithm for the synthesis and training of a hybrid GMDH neural network to forecast the time series.

The following types of neurons will be used for HNN synthesis: classical neuron, nonlinear adaline, R-neuron, W-neuron, wavelet-neuron [6].

Each neuron node in the GMDH network has two inputs. The size of the input window, consisting of a

known sequence of time series can be adjusted. The larger the size of the data window fed to the input of the network – the better the prediction, but the computational cost of the network learning increases, because it is necessary to increase the number of neurons in the input layer. After the formation of the first layer and the initialization of all variables, the first layer is learnt separately according to the training procedure of layer neurons-nodes. After training the first layer, it's selected the n best neurons. In this case, n does not exceed the dimension of the input vector. After selection of the best neurons the new layer having C_2^n neurons is formed. It is necessary to freeze the weights and parameters of the first layer and train the second layer. After training the second layer, we select the best neurons that supply a prediction error less than the best neuron in the previous layer, but not more than $n - 1$ neurons. Repeat the procedure until we can't form a new layer. Then the output of the last layer will be considered the output of the system, and its error – a system error.

In order to simplify the process of structural synthesis of the system and obtain new properties of the model due to hybridity, it is proposed to supplement the above algorithm for constructing an evolutionary GMDH-network by the best type of neurons choice for each layer. That is, to saturate the hybrid GMDH-network with neurons of different types.

To do this, at each iteration of network learning, it is necessary to train several sets of different types neurons. The same input data vector must be applied to the input of each set. After adjusting the network weights, it is necessary to determine the root mean

square error for each of the sets of neurons and select the type of neurons which supply the least root mean square error. After that, it is necessary to discard all sets of neurons, except that it has the least standard error. Next, from the selected neurons it is necessary to choose the best neurons according to the classical algorithm of GMDH and finally form a network layer. The rest of the model is built according to the classical algorithm of GMDH.

With the use of such an algorithm, the need for an expert to decide which neurons will be included in the model disappears and it is possible to obtain new properties and improve the initial accuracy due to the hybridity of the model.

IV. SOFTWARE DEVELOPMENT

A. Software structure

The developed application consists of five subroutines.

- 1) Subroutine of data extraction and division of the training sample.
- 2) Subroutine of data processing and transformation.
- 3) Subroutine of network input vectors formation.
- 4) Subroutine of artificial neural network training and synthesis.
- 5) Subroutine of forecast accuracy check.

All routines except the first can be used for any data set. The first subroutine works directly with data from the source. For the control case, the data on the course of the COVID-19 pandemic, collected by Johns Hopkins University, was used.

B. Reference example

Example of a report of Johns Hopkins University database is represented in a screenshot (Fig. 1).

	FIPS	Admin2	Province_State	Country_Region	Last_Update	Lat	Long_	Confirmed	Deaths	Recovered	Active	Combined_Key
0	45001.0	Abbeville	South Carolina	US	2020-04-05 23:06:45	34.223334	-82.461707	6	0	0	0	Abbeville, South Carolina, US
1	22001.0	Acadia	Louisiana	US	2020-04-05 23:06:45	30.295085	-92.414197	75	2	0	0	Acadia, Louisiana, US
2	51001.0	Accomack	Virginia	US	2020-04-05 23:06:45	37.767072	-75.632346	11	0	0	0	Accomack, Virginia, US
3	16001.0	Ada	Idaho	US	2020-04-05 23:06:45	43.452658	-116.241552	385	3	0	0	Ada, Idaho, US
4	19001.0	Adair	Iowa	US	2020-04-05 23:06:45	41.330756	-94.471059	1	0	0	0	Adair, Iowa, US
...
2758	NaN	NaN	NaN	West Bank and Gaza	2020-04-05 23:06:26	31.952200	35.233200	237	1	25	211	West Bank and Gaza
2759	NaN	NaN	NaN	Western Sahara	2020-04-05 23:06:26	24.215500	-12.885800	4	0	0	4	„Western Sahara
2760	NaN	NaN	NaN	Zambia	2020-04-05 23:06:26	-13.133897	27.849332	39	1	3	35	Zambia
2761	NaN	NaN	NaN	Zimbabwe	2020-04-05 23:06:26	-19.015438	29.154857	9	1	0	8	Zimbabwe
2762	NaN	unassigned	Wyoming	US	2020-04-05 23:06:45	NaN	NaN	0	0	0	0	unassigned, Wyoming, US

2763 rows x 12 columns

Fig. 1. Data on the course of the pandemic in the world for 20.04.2020

The data was downloaded and processed, namely:

- data from different countries were divided. Data from the USA, China, Italy and Ukraine were selected for the experiments;

- separate tables were created for each of the countries, where each row contains information about a certain day;

- added columns with the number of new deaths, new confirmed cases of COVID-19, new recoveries only for this day;

- added a column with the number of active patients as of the current day – added a column with the relative mortality rate, calculated as:

$$d - perc = \frac{\text{new deaths per day}}{\text{number of active sick persons per day}}, \quad (1)$$

- similarly added the percentage increase in new cases and recovered.

At the time of the start of the experiments, 77 days of epidemic surveillance statistics were available for China.

For submission to the model, the data were normalized, and the days when the data were missing were replaced by the average value between the previous and next day. For China, where the epidemic is over, we will chart the active cases of COVID-19 to visually assess the quality of the data (Fig. 2).

The data is visually similar to the normal distribution, but it is clear that at the beginning of the epidemic the quality of reporting was slightly worse, which can not but affect the quality of

prediction. We predict different parameters with different types of neurons and in different modes.

Compared to data from China, reporting from Italy and the United States has fewer data gaps and better reporting accuracy. The worst quality of morbidity data from Ukraine.

```
chinese_data[month][4]
{
  'Active': 2267,
  'Confirmed': 82543,
  'Death': 3330,
  'Recovered': 76946,
  'c_new': 32,
  'c_perc': 0.0141155712395236,
  'd_new': 4,
  'd_perc': 0.00176444640494045,
  'r_new': 186,
  'r_perc': 0.08204675782973092
}
```

Fig. 2. Example of processed data on the epidemic in China in one day

The purpose of the experiment is to determine the quality of forecast for different parameters of the data window and the depth of prediction and the effectiveness of different types of neurons for the task.

The results of the experiment to predict the time series obtained from the data using the GMDH model are represented in Table I. When building and learning models, the parameters of the model window (the size of the input sample, which corresponds to the time series and affects the number of neurons in the first layer of the network), the depth of prediction (how many days before we predict a certain series) and the types of neurons model are changed (Figs 3 and 4).

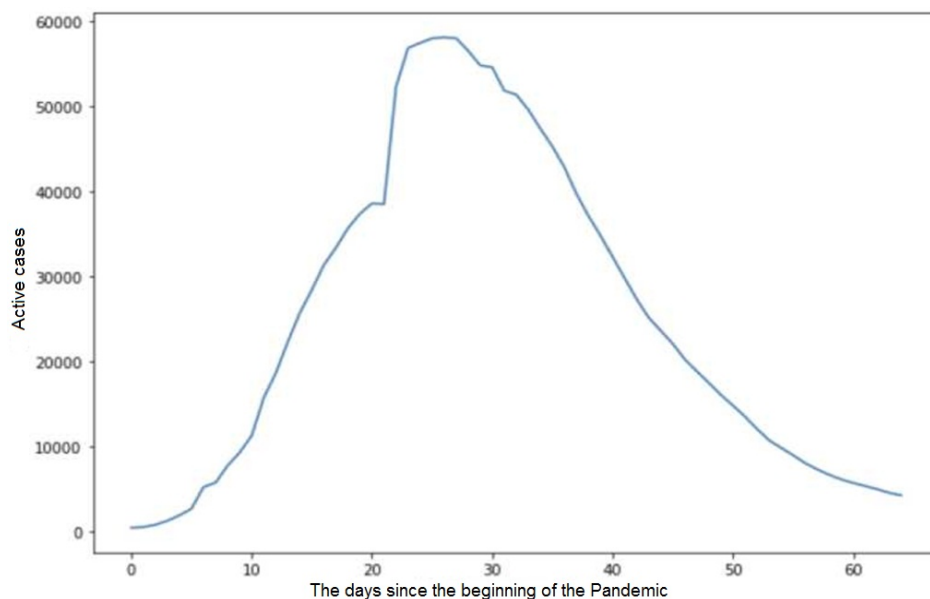


Fig. 3. Active cases of COVID-19 in China

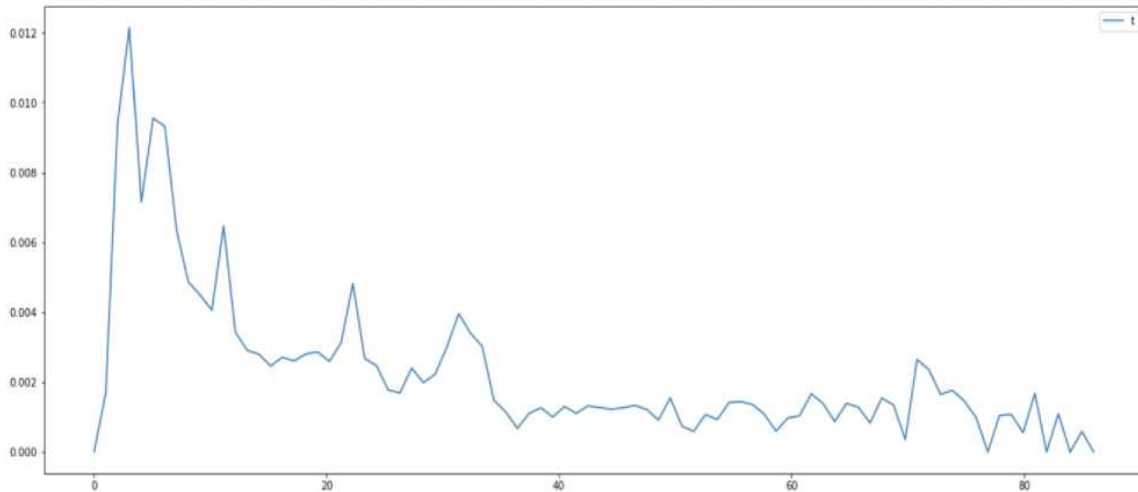


Fig. 4. Graph of relative mortality from Covid-19 in China

TABLE I. RESULTS OF EXPERIMENTS FOR PREDICTION OF PANDEMIC PARAMETERS BY THE CLASSICAL GMDH-NETWORK

No.	Country	Data type	Window	Depth of forecast	Error MAE	Type of neurons
1	China	Relative percentage increase in mortality	5	3	0.0006	<i>Q</i> -neuron
2	China	Relative percentage increase in mortality	5	3	0.00098	<i>W</i> -neuron
3	China	Relative percentage increase in mortality	5	3	0.00081	<i>R</i> -neuron
4	China	Relative percentage increase in mortality	5	3	0.00064	Usual
5	China	Relative percentage increase in mortality	5	3	0.00091	Wavelet-neuron
6	China	Relative percentage increase in mortality	10	1	0.00051	<i>Q</i> -neuron
7	China	Relative percentage increase in mortality	10	6	0.00056	<i>Q</i> -neuron

The models were trained to predict the time series formed from the relative percentage mortality values for each day of the pandemic, calculated by formula (1).

A MAPE error was determined for quality assessment. This prediction quality assessment takes into account only the absolute value of the error because the absolute values of the predictions are very close to zero.

As can be seen from Table I, the increase in the data window has a positive effect on accuracy, but within the available resources does not critically affect the final result. On the contrary, the greater the depth of prediction, the lower the accuracy of prediction. The *Q*-neuron coped best with the task, the worst – *W*-neuron (Fig. 5).

Of interest, of course, is the ability to predict future morbidity or mortality rates from available

data in one country. However, a person can cope with such a task quite well. Graphs are very similar to normal or exponential.

It is more useful to predict the course of an epidemic that is just beginning in one country based on data from another country where the epidemic has already ended or is almost over. Italy and China are well suited for the role of such countries. The Covid-19 epidemic at the time of the experiment in China was over, while in Italy it was in full swing. In addition, these countries have different absolute values of morbidity and mortality rates. Thus, the purpose of the experiment is to understand whether a model on the data of one country is able to capture the parameters of the mortality distribution to make a successful prediction on the data of another country. The experiment used data on the absolute values of deaths per day. The model was constructed using

Q -neurons (Fig. 6). Since the predicted values are much larger than zero, the MARE accuracy criterion was used – the relative average value of the error.

$$MAPE = \frac{1}{N} \sum_{t=1}^N \frac{|Z(t) - \hat{Z}(t)|}{Z(t)} 100\%. \quad (2)$$

where $Z(t)$ is the actual value of Z at time t ; \hat{Z} is forecast value of Z at time t obtained at the previous time; N is the the number of moments (intervals) of time by which the forecast accuracy is estimated.

As can be seen from the table of results, the model, although with some loss of accuracy, was able to successfully apply the patterns learnt from Chinese data to predict the course of the epidemic in Italy.

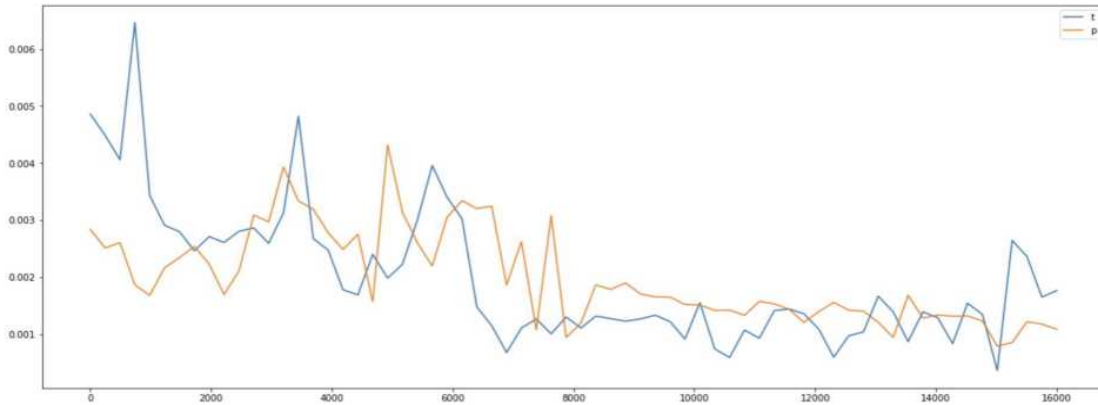


Fig. 5. Results of forecast of NN with wavelet neurons, according to the parameters of window 5 and the depth of prediction 3 to predict the relative mortality in China from Covid-19

From the results (see Table I) it can be seen that to predict time series of different nature and character the neurons can show better or worse results of accuracy relative to each other. Without additional data on the nature of the simulated process, it is impossible to say without experiments which of the neurons is better to use in a particular case. Therefore, the question of network synthesis and selection of neurons for a specific task remains relevant.

We test the proposed algorithm for the synthesis of the hybrid architecture of the GMDH network. The model will be trained to predict mortality from coronavirus COVID-19 in China for the period from January 1 to April 12, 2020. It is expected that the

algorithm will automatically select the network architecture and types of neurons best suited for the task, and the resulting model will have an accuracy no worse than the classical GMDH model using a neuron of the same type as nodes of all layers (Table II) network architecture.

The resulting architecture is slightly more efficient than the architecture built with a single neuron (Table I), but more importantly, completely removed from the human expert the need to determine the most efficient type of neurons and their sequence in the synthesis of a hybrid network to solve the problem of time series forecast. which is obtained by observing processes that have unknown internal dependencies.

TABLE II. EXPERIMENTS OF PANDEMIC PARAMETERS FORECAST IN ITALY BY THE CLASSIC GMDH NETWORK WITH TRAINING ON DATA FROM CHINA

No.	Country	Data type	Window	Depth of forecast	MAPE Error, %	Type of neuron
1	China	New ones died in a day	10	1	27.1	Q -neuron
2	China	New ones died in a day	10	4	30.2	Q -neuron
3	China	New ones died in a day	10	4	35.3	Q -neuron

TABLE III. ACCURACY ANALYSIS OF HYBRID GMDH –MODEL WITH DIFFERENT NEURONS

No. layer	Neuron type	Window	Depth of forecast	MAPE Error, %
1	R -neuron	10	4	30.7
2	Q -neuron	10	4	28.2
3	Q -neuron	10	4	26.5

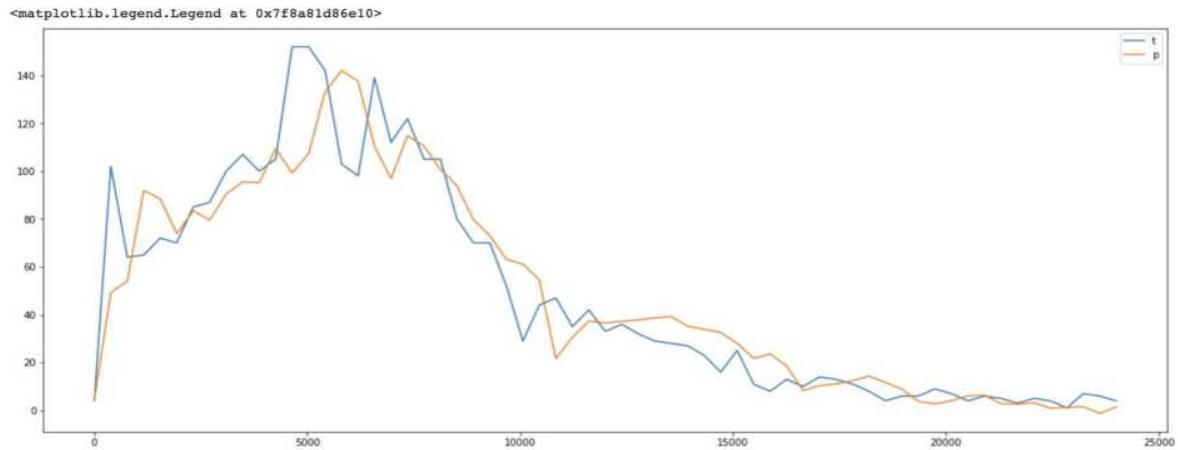


Fig. 6. The model is formed of Q -neurons, trained on data from China to forecast absolute mortality in Italy from Covid-19

V. CONCLUSION

The developed software application allows to synthesize and train GMDH models in different modes and with variable training parameters without interfering with the program code, but with the help of a graphical user interface.

The description of the graphical interface, input and output data of the program and the structure of subroutines are given.

Testing the application on a set of artificial and real data demonstrated the correctness of the synthesis and training of GMDH models and allowed to test the effectiveness of different neuronal topologies as a node of the GMDH network to solve the problem of time series prediction.

Testing of the proposed algorithm for the synthesis of a hybrid network of GMDH showed that the algorithm is able to facilitate the process of architecture synthesis for the human expert without loss of accuracy.

REFERENCES

- [1] V. M. Sineglazov, E. I. Chumachenko, and V. S. Gorbatyuk, *Intelligent methods of forecast*. Kyiv: Osvita Ukraine, 2013, 236 p.
- [2] Ye. Bodyanskiy and O. Vynokurova, "Compartmental adaptive wavelon and its learning algorithm," *Control systems and computers 1*, pp. 47–53, 2009.
- [3] O. I. Chumachenko, S. T. Dychko and A. R. Rizhiy, "Hybrid Neuron Networks Based on Q -, W - and Classical Neurons," *Electronics and Control Systems*, pp. 58–62, 2019. <https://doi.org/10.18372/1990-5548.62.14385>
- [4] Ye. Bodyanskiy, I. Pliss, O. Vynokurova, "Hybrid GMDH-neural network of computational intelligence," *Proc. 3rd International Workshop on Inductive Modelling*, Poland, Krynica. 2009, pp. 100–107.
- [5] Ye. Bodyanskiy, O. Vynokurova, and N. Teslenko, "Cascade GMDH-Wavelet-Neuro-Fuzzy Network," *The 4th International Workshop on Inductive Modelling IWIM 2011*, pp. 22–30.
- [6] E. A. Vinokurova, "Generalized multidimensional wavelet-neuro-fuzzy system in computational intelligence problems," *Intelligent decision-making systems and problems of computational intelligence: a collection of scientific papers based on the materials of an international scientific conference*. Evpatoriya–Cherson, vol. 2, 2010, pp. 329–333.

Received 24 March 2020.

Chumachenko Olena. orcid.org/0000-0003-3006-7460.

Doctor of Engineering Science. Associate Professor.

Technical Cybernetic Department, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute," Kyiv, Ukraine.

Education: Georgian Polytechnic Institute, Tbilisi, Georgia, (1980).

Research area: system analysis, artificial neuron networks.

Publications: more than 80 papers.

E-mail: chumachecko@tk.kpi.ua

Kot Anatoliy. Post-graduate student.

Technical Cybernetic Department, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute," Kyiv, Ukraine.

Education: National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute," Kyiv, Ukraine, (2017).

Research area: artificial Intelligence.

Publications: 4.

E-mail: anatoly.kot@gmail.com

Mandrenko Anja. Bachelor.

Technical Cybernetic Department, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute," Kyiv, Ukraine.

Research area: artificial neural network, deep learning.

Publications: 1.

E-mail: anya.mand@gmail.com

О. І. Чумаченко, А. Т. Кот, Г. Є. Мандренко. Алгоритм побудови гібридної МГУА нейронної мережі для прогнозування часових рядів

Розглянуто проблему структурно-параметричного синтезу гібридних нейронних мереж на основі використання групового методу обробки даних нейронної мережі. Гібридизація досягається за рахунок використання різних нейронів: класичного, нелінійного Адаліну, R-нейрона, W-нейрона, Вейвлета-нейрона. Проблема структурно-параметричного синтезу гібридної нейронної мережі полягає в оптимальному виборі кількості шарів, кількості нейронів у шарах, порядку чергування шарів з різними нейронами. Як приклад розглядається прогнозування рішення проблеми за допомогою гібридних нейронних мереж на основі даних пандемії COVID-19, зібраних університетом Джона Хопкінса. Для оцінювання якості використовувався критерій MAPE.

Ключові слова: гібридна нейронна мережа; структурно-параметричний синтез; проблема оптимізації.

Чумаченко Олена Іллівна. orcid.org/0000-0003-3006-7460. Доктор технічних наук. Доцент.

Кафедра технічної кібернетики, Національний технічний університет України «Київський політехнічний інститут ім. Ігоря Сікорського», Київ, Україна.

Освіта: Грузинський політехнічний інститут, Тбілісі, Грузія, (1980).

Напрямок наукової діяльності: системний аналіз, штучні нейронні мережі.

Кількість публікацій: понад 80 наукових робіт.

E-mail: chumachecko@tk.kpi.ua

Кот Анатолій Тарасович. Аспірант.

Кафедра технічної кібернетики, Національний технічний університет України «Київський політехнічний інститут ім. Ігоря Сікорського», Київ, Україна.

Освіта: Національний технічний університет України «Київський політехнічний інститут ім. Ігоря Сікорського», Київ, Україна, (2017).

Напрямок наукової діяльності: штучний інтелект.

Кількість публікацій: 5.

E-mail: anatoly.kot@gmail.com

Мандренко Ганна Євгенівна. Бакалавр.

Кафедра технічної кібернетики, Національний технічний університет України «Київський політехнічний інститут ім. Ігоря Сікорського», Київ, Україна.

Напрямок наукової діяльності: штучна нейронна мережа, глибоке навчання.

Кількість публікацій: 1.

E-mail: anya.mand@gmail.com

Е. И. Чумаченко, А. Т. Кот, А. Е. Мандренко. Алгоритм построения гибридной МГУА нейронной сети для прогнозирования временных рядов

Рассмотрена проблема структурно-параметрического синтеза гибридных нейронных сетей на основе использования группового метода обработки данных нейронной сети. Гибридизация достигается за счет использования различных нейронов: классических, нелинейных, щелочных, R-нейронов, W-нейронов, вейвлет-нейронов. Задача структурно-параметрического синтеза гибридной нейронной сети заключается в оптимальном выборе количества слоев, количества нейронов в слоях, порядка чередования слоев с разными нейронами. В качестве примера рассмотрено решение задачи прогноза с помощью гибридных нейронных сетей на основе данных пандемии COVID-19, собранной Университетом Джона Хопкинса. Критерий MAPE был использован для оценки качества.

Ключевые слова: гибридная нейронная сеть; структурно-параметрический синтез; задача оптимизации.

Чумаченко Елена Ильинична. orcid.org/0000-0003-3006-7460. Доктор технических наук. Доцент.

Кафедра технической кибернетики, Национальный технический университет Украины «Киевский политехнический институт им. Игоря Сикорского», Киев, Украина.

Образование: Грузинский политехнический институт, Тбилиси, Грузия, (1980).

Направление научной деятельности: системный анализ, искусственные нейронные сети.

Количество публикаций: более 80 научных работ.

E-mail: chumachecko@tk.kpi.ua

Кот Анатолий Тарасович. Аспирант.

Кафедра технической кибернетики, Национальный технический университет Украины «Киевский политехнический институт им. Игоря Сикорского», Киев, Украина.

Образование: Национальный технический университет Украины «Киевский политехнический институт им. Игоря Сикорского», Киев, Украина, (2017).

Направление научной деятельности: искусственный интеллект.

Количество публикаций: 5.

E-mail: anatoly.kot@gmail.com

Мандренко Анна Евгеньевна. Бакалавр.

Кафедра технической кибернетики, Национальный технический университет Украины «Киевский политехнический институт им. Игоря Сикорского», Киев, Украина.

Направление научной деятельности: искусственная нейронная сеть, глубокое обучение.

Количество публикаций: 1.

E-mail: anya.mand@gmail.com