UDC 004.8.032.26; 57.089 **doi: 10.26906/SUNZ.2021.2.053**

S. Aloshyn, I. Khomenko, N. Fursova

National University «Yuri Kondratyuk Poltava Polytechnic», Poltava, Ukraine

COVID-19 CORONAVIRUS SCREENING ANALYSIS NEURAL NETWORK TECHNOLOGY

Abstract. Low-cost, reliable and quick screening diagnosis of coronavirus can be implemented on the basis of intelligent technologies for analyzing a set of signs and symptoms with solving the problem of pattern recognition in the basis of artificial neural networks. The high degree of coronavirus infection diagnostic procedure uncertainty, the vector dimension of input factor-symptoms, fuzzy conditioning and poor formalizability of the subject condition connection with these symptoms require appropriate analytical tools. An analysis of the problem and possible solutions allows justifying the feasibility of implementing screening diagnostics as a solution to the problem of nonlinear optimization in a multidimensional space of high-dimensional factors and states. Artificial neural networks with compulsory training on a representative sample were chosen as a tool for implementing the project. The proposed technology brings diagnostics of coronavirus infection closer to full automation, robotization and intellectualization of complex monitoring (diagnostic) systems as the most promising technology for pattern recognition in systems with a high degree of entropy and allows you to solve the problem at the lowest cost and required performance indicators.

Keywords: neural network, recognition, signs-symptoms, adaptation, modification, pre-processing.

Analysis of the problem

Nowadays, reliable and prompt diagnosis of COVID-19 without hospitalization and quarantine is an urgent and paramount task of society. At the existing technological level of mathematical, algorithmic and instrumental software, there is a possibility of pragmatic and effective introduction of intelligent information and telecommunication technologies at the stages of COVID-19 danger diagnostic procedure (CVD) creation. Particularly interesting variant is the option of mass and rapid diagnostics based on a set of non-invasive signs, the so-called screening - diagnostics.

In this variant there is a fairly high degree of uncertainty in the diagnostic system (high entropy of the system), which logically requires appropriate technology [1]. And this is the technology of data mining in the database of forced learning artificial neural networks [1-4]. With the unique ability to approximate multidimensional continuous functions of any complexity and any ahead given accuracy, under certain conditions, they can be used to find analytical correlations between symptoms and the degree of CVD.

The problem of diagnosing coronavirus danger is that the manifestations of the same symptoms for subjects of different ages, sex, profession, lifestyle, etc. are different. Consequently, the diagnostic system should be based on the scientific and technological basis for the analysis of complex systems, which should be understood as objects with high-dimensional input factors and output states with poorly formalized, unclearly conditioned relationships and a high degree of a priori uncertainty [4]. The high dimension of factors and states, poorly formalized and not clearly determined relationships, the presence of a subjective, human factor, generate a high degree of uncertainty in the diagnostic process, and as a result, lead to unacceptably large errors. It is the feature that does not allow to formalize the diagnosis of CVD in full, to establish analytical dependences of viral danger states with signs-symptoms.

The requirement of the time - full automation, robotization and intellectualization of complex monitoring (diagnostic) systems does not allow the scientific community to evade the introduction of artificial intelligence systems and methods into practice as the most promising technology for pattern recognition in systems with a high degree of entropy. Thus, the efforts of researchers to synthesize an adequate mathematical model of the CV diagnostics process can help to solve the problem of diagnosis reliability in screening programs, which will ensure the use of optimal decision rules and allow the instrumental implementation of the entire complex of analytical and control options in automatic mode [3, 4, 7].

Clear and practice-tested criteria for the quality of pattern recognition and informative factors facilitates to the solvation. Then, using the technology of diagnosing subjects [1-4], it is expedient to concentrate modeling screening analysis on the problem of nonlinear optimization in a multidimensional space of factors and high dimension states [4]. At the same time, it is necessary to select quality criteria in an expert way, to put them in accordance with the values of the states of CVD and accompanying symptoms using examples of the existing prehistory (retrospective data), and strict analytical connections must be established and fixed by forced training of neural network models on a retrospective sample from real practice [1, 3].

Thus, the subject expert describes the task at the verbal level in terms of the medico-indicative paradigm, and a specialist in information technologies formalizes the task and its instrumental solution.

As it is showed in [1, 4, 6] for objects of the type under study, it is advisable to apply intelligent data analysis procedures in the basis of artificial neural networks.

This choice of analytical tools is due to the indisputable and important for practical use advantages:

- neural networks are adaptive learning systems that use examples to extract information from real processes that are difficult to dynamically simulate, because they often contain a significant array of hidden, uncontrolled, incomplete and noisy parameters and interconnections between them;

- their application allows solving problems that are difficult or impossible to solve by traditional methods due to the absence of formalized mathematical descriptions of the functioning processes;

- have associative memory and in the process of work accumulate and generalize information, from which their effectiveness increases over time;

- their use is based on training a neural network to extract information from empirical data, which ensures the objectivity of the results and increases their reliability and validity.

At the same time, modern software tools allow to design a neural network environment and implement a wide class of neural network architectures of varying complexity and rules for modifying weight coefficients in the process of forced learning with the ability to adapt a set of standard options (preprocessing, factor analysis, organization of homogeneous subsets, input sensitivity, classification, etc.)

The extent of problem study

The development, theoretical justification and introduction of information technologies to solve the problems of image recognition in the neural network paradigm were engaged by national and foreign scientists, in particular: R. Hecht-Nielsen, T. Kohonen, S. Haikin, etc.

As a result of their activities there are:

Fundamental principles and methods of applied artificial intelligence theory in social and technical systems;

Constructive approaches to designing and modeling complex social and production facilities;

Methods and algorithms of multi-level differential diagnostics in various subject areas;

Algorithmic and software products for rapid multifactorial data analysis in the tasks of classification, forecasting and adapting inputs to the states of the object.

At the same time, the comprehensive and effective software base of screening diagnostics automation in the space of noisy, low-informative signs of high dimension is not widely advertised in open publications today. And the critical situation with the pandemic requires today the immediate creation of productive technologies and software, the construction of reliable diagnostic models, allowing to automate the processes of CVD recognition by establishing determinized connections of criteria, states, symptoms based on intelligent technologies.

Purpose of the article

Develop the technology of automatic screening-analysis neural network model productive ensemble synthesis in space of not invasive signs - symptoms for the operational, reliable and mass monitoring of CVD state of society and its subjects.

Setting a task. It is necessary to describe the state of CVD as a class in the language of its informative features, to find an analytical display of the data input vector on the result of the assessment of subject CVD state.

Building a mathematical model of CVO state indicator internal structure based on input data is down to mapping the factors into the space of states with specified reliability and accuracy:

$$F : X \to Y_{opt}, X \subset \Re^m, Y_{opt} \subset \Re \ , \quad (1)$$

where X is a vector of signs of the state of the CVD;

Y_{opt} is the output value of the CVD state class.

An array of CVD traits, in conjunction with the relevant array of CVD state classes, allows to implement a known $X^n = \{x_1, x_2, \ldots, x_n\} \subset X$ pattern recognition rule [6].

$$\omega_g \in \Omega_k \text{ If } L(\omega, \{\omega_g\}) \sup_i L(\omega, \{\omega_i\}) \quad (2)$$

$$L(\omega, \{\omega_g\}) \to \omega_g \in \Omega_k \, ,$$

where $\vec{X} = (x_{1,\ldots\ldots,}x_n) \in X$,

$L(\omega, \{\omega_g\})$ - the rule of categorizing CVD state ω_g to the appropriate class;

$\{\omega\}$ - many states of the CVD *(p, g)* in the space of the signs*(k,l)* for all their possible combinations (ω_{pk} , ω_{gl}) .

Approximation of current CVD state function

The productive application of the neural network approach to the analysis of CVD relies on the Kolmogorov-Arnold theorem of presenting the function of several arguments through the sum of functions of one variable and its adaptation to the Hecht-Nielsen neural network format. Then the vector *y* (CVD classes) and the associated symptoms-signs of the current state of the subject can be presented as the following:

$$y(x) = \alpha \sum_{i=1}^{H} v_i \left(w_{i1}x_1 + w_{i2}x_2 + \ldots + w_{in}x_n + u_i \right), \quad (3)$$

where H is the power of the training sample, the parameters of the neural network is α, v_i, n – he number of neurons, w_{i1}, w_{i2}, …, w_{in} is the weight ratios of neurons. At the same time, it can be argued that there is a set of numbers H, n, α, v_i, u_i for which the function is approximated by series (3) over the entire domain of its definition and can be implemented using a three-layer neural network with any predetermined error. This is a fundamental position, for it is the basis of all subsequent procedures. Indeed, if it is possible to establish a deterministic analytical relationship of the entire set of symptoms with CVD level (class), then all subsequent options are derived from this key procedure. Therefore, it requires a lot of effort and time.

Formation of a training sample

Nowadays, not invasive signs of COVID-19 are recorded and statistically reliable (recent data in open reports of subject specialists): headache; temperature of the body; muscle pain; abdominal pain; drowsiness; local pain; lack of appetite; pulse; receptor sensitivity, %; irritability.

In a simple version the function of CVO state can be described in two classes: 1 (Healthy); 2 (Ill).

Training and modification of neural network synaptic set. The neural network training procedure is based on the over-reassession of all combinations of synapses, for example, on the algorithm of reverse distribution of the error [1-3] and implemented by the neural network module of the technical analysis package:

$$w_{hq}^{(n)}(t) = w_q(t-1) + \Delta w_{hq}^{(n)}(t), \qquad (4)$$

where $w_q(t-1) = w_q(t) + \alpha \cdot \dfrac{\partial E(k)}{\partial w_q(t)}$,

w - the array of synaptic coefficients;

q - the number of neuron output in the -n layer;

h - neuron entry number in -n layer;

n - network layer number

The process of modifying weights in the training of the neural network continues until the condition is met:

$$\frac{1}{mn}\sum_{j=1}^{n}\sum_{i=1}^{m}(y_{ij} - d_{ij})^2 \Rightarrow \min\ (\Re \le \Re_0), \qquad (5)$$

y_{ij} - current state of the CVD;

d_{ij} - the result of the network's learning (response) on j - output, with i - m example of a training sample;

$j = 1, n$ - the network's exit number;

$i = 1, m$ - the example number;

m, n - size of the sample array and the number of network element outputs.

Instrumental modeling of neural network model ensemble

In order to solve the problem, standard technical analysis package was used: table of original data (200-300 lines of training sample) and the program online.

Firstly, launch the Statistica Neural Networks module. In the starting window choose the type of task: "Classification" and "Master of Solutions", set the function and arguments on the working tabs. Choose the number of models to be saved (e.g.5). Then choose the ratio of sampling types, conduct training, evaluate the result. On the tab "Fast", select the "Network Summary" option and find the best one in terms of performance. The set of options is simple and accessible to a wide range of users, which contributes to the widespread implementation of the proposed technology into practice.

Since each class corresponds to its own set of indicators, the procedure of classifying the object is reduced to the analysis of current state signs and comparing the results of the analysis with descriptions of early chosen classes. In the neural network format, the problem of investigated subject current state class recognizing is solved, for example, using the delta rule [3]. In the presence of two classes of states, the formation of the training set is simplified. Then it is advisable to apply the modeling technology in the environment of neuroemulators according to the method [1-3].

Lower performance is typical of models with a large number of neurons in a hidden layer, and it is necessary to find a trade-off between generalization errors on the test set, learning time, and network performance. Therefore, the choice of the final model is made by the user based on the features of subject under study, the requirements for models and the capabilities of the used neuromedia (Fig. 1.)

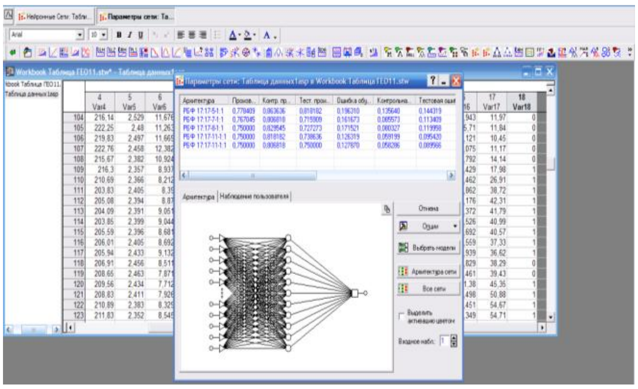

Fig. 1. Model profiles on a sample fragment

And under certain conditions and other paradigms of neural networks, activation functions, training methods, dimensions of the input symptoms, the power of the training sample, etc. performance indicators can be significantly improved (Fig. 2).

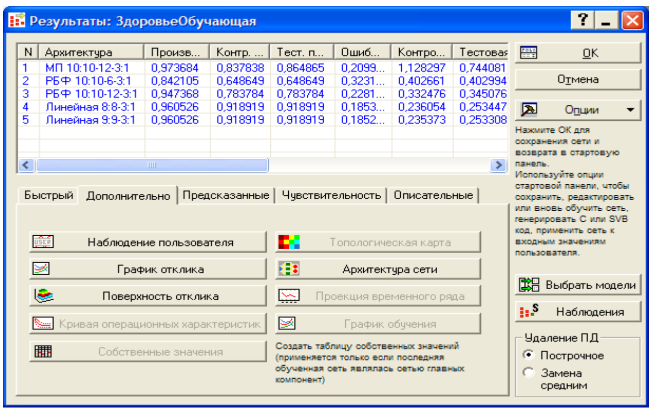

Fig. 2. Ensemble of different paradigm models of MPL, RBF, Linear Network

Model performance depends on the power of the sample at the same time, the complexity of the network, its type of and the architecture. Thus, there is a theoretical and practical confirmation of the conclusion that it is possible to build a model of CVD while ensuring the adequacy of models within the specified boundaries. The practical implementation of the project requires the most informative features of CVD, the greater statistical validity of the training sample from the data of pandemic coronavirus real dynamics.

If the space of CVD features is supplemented by the results of the detailed biochemical analysis of blood, tomography, X-ray, etc., the proposed screening analysis technology can be transformed into an intelligent application into a classical system of complete laboratory diagnostics of CVD. Only the dimension of the input vector, the power of the training set and the time of result achievement will change. However, even qualitative analysis of the convergence of the iterative process in different variants of data and learning conditions demonstrates its steady dynamics, which is shown in the two graphs of the MPL model training with different architectures (Fig. 3, 4).

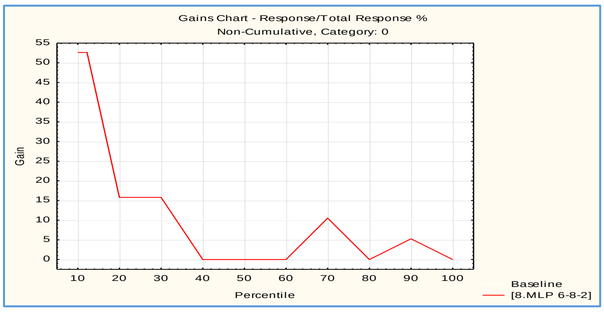

Fig. 3. MPL model training (Option 1)

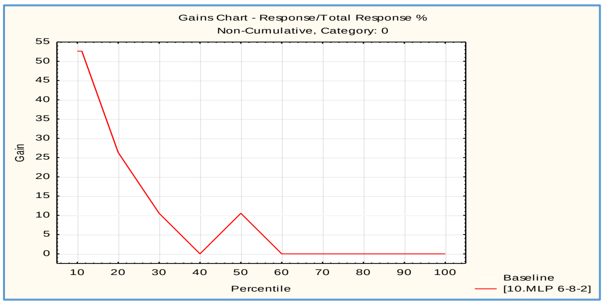

Fig. 4. MPL model training (Option 2)

As it can be seen from the graphs, the application of data mining technology in the form of neural network forced training in the space of input symptoms-factors, even in conditions of high prior uncertainty in the task of CVD classification, allowed to obtain a stable convergence of synaptic space modification iterative process (Fig. 3, 4). The number of epochs does not exceed a few hundred, which in terms of total time costs corresponds to units and tens of seconds.

Thus, the analysis of simulation the results in the neural network format generally confirms the theoretical conclusion about the productivity of CVD model analysis in real time. The results of the simulation in the basis of various structural network constructions, training methods and variations in the dimension of input symptoms-factors, as well as the power of the training set, showed a steady convergence of the iterative process of training networks with acceptable performance indicators.

For different modeling conditions, an ensemble of productive neural networks has been obtained that can be used in individual applications based on their preservation in the main code format.

Justification of simulation result consistency

When determining a classification error, the result of an erroneous decision is considered an event when an object belonging to the i- class is classified as an object of the j- class. In our case, in the study of CVD, the same signs of different classes take on their values at coinciding intervals, so errors are inevitable. It is necessary to minimize errors in this situation.

When the alphabet of classes is compiled, that is, for each object b_j, a set of attributes $\{x_i\}_j$ is assigned and the object is completely determined by them. Analysis of the description of the state of the object gives current information, on the basis of which the posteriori information is determined. The nature of the current and a posteriori information is determined by the decision rule implemented in the classifier. Since the ultimate goal is to maximize the reliability of decisions, it is advisable to apply the classical statistical rule of the ideal observer. Under these conditions, the definition of current information is reduced to calculating the likelihood function, $P(x_j \setminus A_i)$, i = 1, 2, ..., M, and posteriori - to calculating posterior probabilities [4].

$$h_{ij} = P(A_i \setminus x_j) = \frac{P(A_i)P(x_j \setminus A_i)}{\sum_{i=1}^{M} P(A_i)P(x_j \setminus A_i)}, i = 1,2,\dots,M, \quad (6)$$

where $\sum_{i=1}^{M} P(A_i)P(x_j \setminus A_i)$ is a generalized distribution of class traits (states) probabilities of the object under study (CVD).

The conclusion about the productivity of such an approach is confirmed by the internal generality of decision-making procedures according to the statistical rule, for example, maximizing the posterior probability of CVD class and modifying the synaptic space of an artificial neural network [4]. The base for extracting knowledge about CVD states of the investigated subject in both statistical and neural network formats is the same - an array of precedents, which is an objective condition for the validity of the conclusions made.

Thus, these conclusions apply to each model from our ensemble of trained networks, the productivity of which (the probability of recognizing the state of CVD) is characterized by an individual profile. The practical implementation in the neural network version of the rule for the maximum a posteriori probability of CVD class when testing statistical hypotheses is carried out in the environment of standard neuroemulators Statistica Neural Network [1–3].

For practical problems of recognizing CVD levels, achieving maximum network performance (the probability of class recognition) with admissible error values on the training, control and test sets is a necessary and sufficient condition for ensuring the reliability of solving classification problems [4].

The adequacy of neural network models established by the value of model ensemble productivity, errors on training, control and test sets when training on a representative sample allows to assert the consistency of decisions made based on the results of modeling [1, 4, 7]. In practice, it allows to respond to possible changes quickly in the dynamics of CVD symptoms by making changes to the original training set and carrying out the procedure for retraining neural networks in order to keep the efficiency criteria without significant material and time costs.

Conclusions

1. To automate the determination of CVD class, it is necessary to find the functional dependence of its states on the values of factors. This problem is solved by using the technology of neural network forced learning and implemented by models of multilayer perceptrons as a recognition problem.

2. The practical significance of the research results lies in the creation of software tools for the transition to automatic systems of CVD screening analysis in the space of input features.

3. The developed technology, methodological, algorithmic and software tools allow automating the processes of classifying CVD states, to minimize the influence of the human factor.

4. Modeling and interpretation of the results on the platform of neuroemulators is an effective means and tool for automating decision-making management in real time for monitoring CVD.

5. Functionally, the program of trained models can be implemented as an independent application in the main code of technical data analysis package in the format of the basic subsystem for classifying the states of the research object.

6. Automation of CVD monitoring, implemented on the basis of the practical use of artificial intelligence capabilities in a neural network format, makes it possible to increase the productivity of diagnostics, reduce the time for making a decision and the risks of a pandemic spread

REFERENCES

1. Haikin, S. Neural Networks: Full course. 2nd edition / Haikin S.M.: "Williams," 2006. 1104 s.
2. Galushkin A.I. Neurocomputers and their use at the turn of the millennium in China. T.1 and 2 / A.I. Galushkin. M., 2004. 367 p., 464 p.
3. Gorban A.N., Rossiev D.A. Neural Networks on a personal computer / Novosibirsk: Science, 1996. 276 p.
4. Alyoshin S.P. Neuronet base of support solutions in the space of factors and states of high dimension / Monograph - Poltava: Izd. Skytech, 2013. 208 p.
5. Neuronet control of the dynamics of processes as space states of high dimension / S.P. Alyoshin, E.A. Borodina // Herald of RGUPS. 2013. № 4. P. 35 - 42.
6. Neuronet recognition of classes in real time (Electronic resource) / S.P. Alyoshin, E.A. Borodina // Don's Engineer Gazette. 2013. № 1. Access mode: http://www. ivdon.ru/magazine/archive/n1y 2013/1494.
7. Neuronet modification of the current space features to the target set of classes / A.L. Lyakhov, S.P. Alyoshin, E.A. Borodina // Vysnik Donbass power machine-first academy. 2012. № 4 (29). P. 99 – 104

Received (Надійшла) 22.03.2021
Accepted for publication (Прийнята до друку) 19.05.2021

Нейромережна технологія підтримки скринінг-аналізу коронавіруса COVID-19

С. П. Альошин, І. В. Хоменко, Н. А. Фурсова

Анотація. Низькозатратна, надійна і оперативна скринінг-діагностика коронавіруса може бути реалізована на основі інтелектуальних технологій аналізу сукупності ознак-симптомів з рішенням задачі розпізнавання образів в базисі штучних нейронних мереж. Високий ступінь апріорної невизначеності діагностичної процедури коронавірусної небезпеки, розмірність вектору вхідних факторів-симптомів, нечітка обумовленість і слабка формалізованість взаємозв'язків станів суб'єкта з цими симптомами вимагають відповідного аналітичного інструментарію. Аналіз проблеми і можливих рішень дозволяє обґрунтувати доцільність реалізації скринінг-діагностування як рішення задачі нелінійної оптимізації в багатовимірному просторі факторів і станів високої розмірності. Як інструмент реалізації проекту обрано штучні нейронні мережі з примусовим навчанням на репрезентативній вибірці прикладів. Проведені дослідження ознакових просторів процесу розпізнавання (діагностування), можливостей коректного застосування статистичних вирішальних правил, алгоритмів примусового навчання ансамблю синтезованих нейромережних моделей в базисі існуючих пакетів технічних даних, дозволяють підвищити продуктивність технічних засобів діагностики шляхом автоматизації процесу аналізу, зниження впливу суб'єктивних рішень, скорочення часу реакції. Запропонована технологія наближає діагностику коронавірусної небезпеки до повної автоматизації, роботизації і інтелектуалізації складних моніторингових (діагностичних) систем як найбільш перспективної технології розпізнавання образів в системах з високим ступенем ентропії і дозволяє вирішувати вартісну задачу при мінімальних витратах і необхідних показниках ефективності. При цьому сучасний програмний інструментарій дозволяє оперативно конструювати нейромережному середу і реалізувати широкий клас архітектур нейромереж різної складності і правил модифікації вагових коефіцієнтів в процесі примусового навчання з можливістю адаптації набору стандартних опцій (попередня обробка, факторний аналіз, організація однорідних підмножин, чутливість входів, класифікація та ін.).

Ключові слова: нейронна мережа, розпізнавання, ознаки-симптоми, адаптація, модифікація, предобробка.