The Methods for Training Technological Multilayered Neural Network Structures

Andrey Kupin[1], Yuriy Osadchuk[2], Rodion Ivchenko[3] and Oleg Gradovoy[4]

[1,2,3,4]*Kryvyi Rih National University, Ukraine, 50027, Kryvyi Rih, Vitaly Matusevich, 11*

Abstract

The analysis of existing methods of training multilayered of neural networks structures is made. The way computer modelling investigates the most effective methods of training. Recommendations of application of the selected methods on an example of problems of multilayered approximation for concentrating technology are given.

Keywords

Multilayered neural networks, training methods, Conjugate gradient, approximation, classification.

1. Introduction

Now even more often to the decision of applied problems of information and automation in the conditions of difficult manufactures apply different technologies of intellectual control [1]. Thus one of base approaches for construction of mathematical models in the course of approximation, identification, classifications are applications of multilayered neural networks (NN) of different architecture.

The complex technological processes of the mining and metallurgical industry are a good basis for demonstrating the benefits of using intelligent technologies. All this is confirmed by the presence of multifactoriality, incomplete information, nonlinear characteristics, nonstationarity, etc. All this provides quite good prerequisites for the successful application of computational intelligence technologies in the automation of basic processes.

For today in the theory of artificial neural networks there are no definite answers on concrete questions of an unequivocal choice of this or that architecture and the most effective method of training (parameterization). Therefore the majority of researchers operate in the empirical image, selecting from certain set of potentially possible alternatives of the best a variant behind certain criteria and in the conditions of concrete technology.

2. The analysis of last researches, publications and problem statement

In numerous works of the authors [2-4], the outstanding capabilities of neural networks and fuzzy logic have been proved for solving problems of automated control of similar technological processes. This is the mining, metallurgical, coal, energy and other industries of Ukraine. However, there are still many unsolved problems associated with various applied aspects of the use of intellectual control. This is, first of all, the choice of the architecture of neural networks, parameter settings, the choice of effective learning algorithms, etc. The metadology of scientific approaches in such conditions can

ICTERI-2021, Vol I: Main Conference, PhD Symposium, Posters and Demonstrations, September 28 – October 2, 2021, Kherson, Ukraine
EMAIL: kupin.andrew@gmail.com (A. Kupin); u.osadchuk@knu.edu.ua (Y. Osadchuk); ivchenko.ra@gmail.com (R. Ivchenko); queke888@gmail.com (O. Gradovoy)
ORCID: 0000-0001-7569-1721 (A. Kupin); 0000-0001-6110-9534 (Y. Osadchuk); 0000-0003-4252-4825 (R. Ivchenko); 0000-0001-6984-1690 (O. Gradovoy)
© 2020 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

differ significantly. This work is intended to solve one of these problems. This is the problem of training neural structures for technological purposes.

For training (parameterization) multilayered of neural network the structures intended for the further identification and control by difficult technological processes (TP) in a mode of real time, it is necessary to apply methods which meet certain requirements. According to [2] these requirements, first of all, concern: speed of convergence, computing robust, requirements concerning an operative computer memory, etc. Among existing methods in the biggest measure respond for today these requirements so-called methods of 2nd order, namely [2-6]:

- Levenberg-Markuardt (LM);
- Gauss-Newton (GN);
- Conjugate Gradient (CG);
- Modifications to these methods.

Therefore the further analysis, research and a choice potentially the most effective methods of training to neural structures of the technological appointment, offered in [1], will be limited to set of these methods. Thus very important from the point of view of automation of the further calculations and modelling is that the specified methods are realised as a part of the most powerful packages of applied programs on emulation of neural network structures (MATLAB Neural Tools, Neuro Solutions, Statistical Neural Network, etc.) [5, 6].

3. Statement of a material and results

All aforementioned methods are based on decomposition functions abreast Taylor to 2nd order inclusive. Such decomposition near to a point (a theoretical optimum of parameters of NN) will have such appearance [4]:

$$\begin{aligned} V_M\{\Theta,S,\Xi\} &= V_M\{\Theta^*,S,\Xi\} + (\Theta-\Theta^*)^T V_M'\{\Theta^*,S,\Xi\} + \\ &+ \frac{1}{2}(\Theta-\Theta^*)^T V_M''\{\Theta^*,S,\Xi\}(\Theta-\Theta^*) = V_M\{\Theta^*,S,\Xi\} + \\ &+ (\Theta-\Theta^*)^T G(\Theta^*) + \frac{1}{2}(\Theta-\Theta^*)^T H(\Theta^*)(\Theta-\Theta^*), \end{aligned} \tag{1}$$

where $V_M\{\cdot\}$ is a designation of criterion of criterion function; Θ is a vector of parameters which are subject to adjustment (architecture of NN, weight factors, depth of regress); S is versions regressive models which it is applied; Ξ is statistical sample of the data for training; $G(\Theta^*)$, $H(\Theta^*)$ are accordingly a gradient and gessian in an optimum point.

The gradient is defined as

$$G(\Theta^*) = V_M'\{\Theta^*,S,\Xi\} = \left.\frac{dV_M\{\Theta^*,S,\Xi\}}{d\Theta}\right|_{\Theta=\Theta^*} \tag{2}$$

and a matrix of the second derivatives – gessian or a matrix Gess

$$H(\Theta^*) = V_M''\{\Theta^*,S,\Xi\} = \left.\frac{d^2V_M\{\Theta^*,S,\Xi\}}{d\Theta^2}\right|_{\Theta=\Theta^*} \tag{3}$$

Null value of a gradient and positive definiteness gessian is sufficient conditions of a minimum of function. That is

$$\begin{cases} G(\Theta^*) = 0 \\ H(\Theta^*) > 0 \end{cases}.$$

In most cases minimum search can be shown to iterative procedure of type:

$$\Theta^{(i+1)} = \Theta^{(i)} + \mu^{(i)} f^{(i)},$$

where $\Theta^{(i)}$ is value of parameters of current iteration (); $f^{(i)}$ is a search direction; $\mu^{(i)}$ is a step of algorithm of current iteration.

Linear approximation of an error of forecasting $\varepsilon(t,\Theta)$ according to an initial signal on the NN exit in $d\hat{y}(t\,|\,\Theta)$ such kind is thus applied $\varepsilon(t,\Theta)$ $d\hat{y}(t\,|\,\Theta)$:

$$\tilde{\varepsilon}(t,\Theta) = \varepsilon(t,\Theta^{(i)}) + (\varepsilon'(t,\Theta^{(i)}))^T(\Theta - \Theta^{(i)}) =$$
$$= \varepsilon(t,\Theta^{(i)}) - (\psi(t,\Theta^{(i)}))^T(\Theta - \Theta^{(i)})^T,$$

where $\psi(t,\Theta) = \dfrac{d\hat{y}(t\,|\,\Theta)}{d\Theta}$, t is value of discrete time.

The modified criterion (1) is:

$$V_M\{\Theta, S, \Xi\} \approx L^{(i)}(\Theta) = \frac{1}{2M}\sum_{i=1}^{M}[\tilde{\varepsilon}(t,\Theta)]^2 ,$$

where $L^{(i)}(\Theta)$ is the approached value of the modified criterion; M is quantity of templates of training sample.

Search direction in Newton - Gauss method it is based on definition of approximation of criterion $L^{(i)}(\Theta)$ around current iteration [2-5]. In turn a method of the interfaced gradients based on change of directions of search (restart) in a direction to a gradient (anti gradient) in the conditions of sharp delay of convergence. Thus there are different lines of thought and algorithms of realisation of the specified procedures for both methods (set of versions [7]).

At the same time in one algorithm it is not considered that the global minimum $L^{(i)}(\Theta)$ can be out of a zone of current iteration therefore search will be incorrect. Therefore more rational will estimate at first expediency of search of a minimum $L^{(i)}(\Theta)$ in the field of current iteration. For this purpose behind algorithm of a method of Levenberg-Markuardts (known in the literature under synonyms: Levenberg-Marquardt methods, the scheme of Levenberg) is selected sphere to radius $\delta^{(i)}$. Then the optimisation problem can be formulated in the form of such system

$$\begin{cases} \hat{\Theta} = \arg\min L^{(i)} \\ \left|\Theta - \Theta^{(i)}\right| \le \delta^{(i)} \end{cases}. \tag{4}$$

Interactive procedure of search of a minimum behind presence of restrictions contains such stages in system

$$\begin{cases} \Theta^{(i+1)} = \Theta^{(i)} + f^{(i)} \\ \left[R(\Theta^{(i)}) + \lambda^{(i)} I\right] f^{(i)} = -G(\Theta^{(i)}) \end{cases}, \tag{5}$$

where $\lambda^{(i)}$ is parameter which defines area $\delta^{(i)}$.

The hypersphere to radius $\delta^{(i)}$ is interpreted as area in which limits $L^{(i)}(\Theta)$ it can be considered as adequate approximation of criterion $V_M\{\Theta, S, \Xi\}$.

Feature of a method is procedure of definition of interrelation between and $\delta^{(i)}$ parameter $\lambda^{(i)}$. As unequivocal dependence between them does not exist, in practice apply some heuristic procedures [2]. For example, the gradual increase $\lambda^{(i)}$ until will take place criterion reduction $L^{(i)}(\Theta)$ then iteration comes to the end. Value of parameter $\lambda^{(i+1)}$ for the following operation decreases.

Also the alternative approach based on comparison of real reduction of criterion and reduction which it is predicted on the basis of approximation is applied $L^{(i)}(\Theta)$. As a measure of accuracy of approximation the factor is considered

$$r^{(i)} = \frac{V_M\{\Theta^{(i)}, S, \Xi\} - V_M\{\Theta^{(i)} + f^{(i)}, S, \Xi\}}{V_M\{\Theta^{(i)}, S, \Xi\} - L^{(i)}(\Theta^{(i)} + f^{(i)})}. \tag{6}$$

In case of approach of value to factor $r^{(i)}$ to 1, $L^{(i)}(\Theta)$ is adequate approximation $V_M\{\Theta, S, \Xi\}$ and value λ decreases that responds increase $\delta^{(i)}$. On the other hand, small or negative values of factor lead to necessity of increase λ. On the basis of it the general scheme of realisation of algorithm the such:

1. To select initial values of a vector of parameters which are subject to adjustment Θ (0), and factor λ (0).
2. To define a direction of search from system of the equations (5).
3. If $r^{(i)} > 0{,}75 \;\Rightarrow\; \lambda^{(i)} = \lambda^{(i)}/2$.

4. If $r^{(i)} < 0{,}25 \quad \Rightarrow \quad \lambda^{(i)} = 2\lambda^{(i)}$.

5. If to accept for new iteration and $\Theta^{(i+1)} = \Theta^{(i)} + f^{(i)}$, then to establish $\lambda^{(i+1)} = \lambda^{(i)}$.

6. If criterion of a stop will not reach, then to pass to a stage 2.

Value of criterion which is minimised, can be presented in such kind

$$L^{(i)}(\Theta^{(i)} + f) = V_M\{\Theta^{(i)}, S, \Xi\} + f^T G(\Theta^{(i)}) + \frac{1}{2} f^T R(\Theta^{(i)}) f \ . \tag{7}$$

Substituting to (2) value of expression for direction finding of search which is received from a parity

$$R(\Theta^{(i)}) f^{(i)} = -G(\Theta^{(i)}) - \lambda f^{(i)},$$

and we set

$$V_M\{\Theta^{(i)}, S, \Xi\} - L^{(i)}(\Theta^{(i)} + f^{(i)}) = \frac{1}{2}\left(-(f^{(i)})^T G(\Theta^{(i)}) + \lambda^{(i)} \mid f^{(i)} \mid^2\right). \tag{8}$$

The parity (8) allows at stages 3, 4 algorithms to define factor on $r^{(i)}$ expression (6).

On the basis of the general technique intellectual neural multilayered identification [8] with application of methods of computer modelling researches of modelling structures on a basis of neural network autoregressive predictors for conditions ТП of concentration quartzites of magnetite have been conducted. Research was included by such stages:

- choice of a method of training, estimation of depth of regress (quantity of the detained signals on an input and an exit) models;
- application of methods of training (speed of convergence, accuracy);
- direct and return forecasting;
- testing of the received systems for nonlinearity.

The analysis and choice of a base set of methods of training for identification models was carried out on the basis of a technique stated in [2]. The basic investigation phases are such:

1. For imitating experiments the elementary model of type NNARX (Neural Network based Autoregressive exogenous signal) has been selected. For the purpose of analysis simplification identical depth of regress ($l_1 = l_2 = 2$) on the basis of the previous results [1, 8] has been accepted $l_1 = l_2 = 2$.

2. Templates of NN of modelling structures in bases of NN of direct distribution (NNDD), radial-basic functions (RBF) that full the coherent (FCNN, recurrent) are prepared. For all models was applied in the NN from one latent layer behind the formula: 16-8-8 (corresponding quantity neurons on a structure input, in the latent layer and on an exit).

3. Tenfold training and testing of all specified NNS of structures with application of four methods of training has been carried out: return distribution of an error (back propagation or BP – a method, as the actual standard from NN training [2-6]), Gauss-Newton (GN - method), Levenberg-Markuardt (LM) and Conjugate gradient (CG). Statistical sample of indicators has been applied to training Northern Mining Complex (Kryviy Rih, Ukraine) behind the formula: 350-280-70 (total of templates, quantity of templates for training, quantity of templates for verification). Base indicators of first and last stage TP were thus analyzed.

4. Average indicators of convergence (the quantity of epoch or iterations for training), robust (a root-mean-square error – MSE, the generalised root-mean-square error -- NMSE [6]) and the applied computing resources (operative memory) has been brought to Table 1.

5. On the basis of the results received in the course of research there was their carried out comparative analysis.

The authors tested other neural network architectures using a similar methodology. Absolutely all research data showed quite encouraging results. Thus, this proves that the approach is quite promising.

Further research will consider more complex neural network architectures based on deep learning. In our opinion, deep neural networks (DNN [7]) may be of the greatest interest for this:

- convolution (CNN),
- recurrent (RNN),
- long short-term memory (LSTM),

- neural networks with an attention mechanism (NNAM).

Convolutional neural networks that use convolutional layers, union layers, fully connected and lost layers to simulate parallel computing. The convolutional layer basically counts the integrals of many small overlapping areas. In a fully connected layer, neurons have connections to all activations in the previous layer. Loss rate calculates how network learning corrects the variance between predicted and true labels using the softmax function or the cross-entropy loss function for classification, or the Euclidean loss function for regression. A network with long short-term memory is capable of forgetting or remembering previous information. LSTM can handle sequences of hundreds of past inputs. Attention modules are generalized elements that apply weights to a vector of inputs. A hierarchical neural attention encoder uses multiple layers of attention modules to work with tens of thousands of past inputs.

Table 1

Comparative estimation of accuracy, resources consumption and speed of convergence of potential algorithms of training investigated neural structures

Algorithm of training	Convergence, Epoch (iterations)	MSE	NMSE	COMPUTER resources, Mb
1. Basis NN (multilayered perceptron)				
1.1. BP	568	1,198596	1,76165223	30
1.2. GN	303	1,161828	1,96306745	24
1.3. LM	177	0,778172	1,45139743	35
1.4. CG	425	0,888760	1,45448391	21
2. Basis RBF (radial-basic functions)				
2.1. BP	196	1,85732511	2,111487478	30
2.2. GN	65	1,19651332	2,131730124	25
2.3. LM	31	0,79076953	1,906790835	35
2.4. CG	87	0,89815021	1,912728683	21
3. Basis FCNN (full coherent neural networks)				
3.1. BP	837	1,0915434	1,60226771	33
3.2. GN	451	1,0807423	1,77265223	27
3.3. LM	265	0,7223413	1,21234453	37
3.4. CG	637	0,8684867	1,26644234	22

As program environments for computer modelling there were applied three independent packages of applied programs (neural simulator) type: Neuro Solution, Statistica Neural Networks and MATLAB Neural Networks Tools (NNT). Corresponding results of modelling in these different packages approximately coincide. Also all received results well enough coincide with resulted in [1, 2].

In the course of computer modelling it has been applied such system hardware-software platform:

- Personal computer with working parameters CPU Pentium IV 2.66 GHz/RAM 8 Gb;
- Operating system MS Windows 10.

On Figure 1 curves which show change of criterion of root-mean-square error MSE (Mean-Square Error or Normalized NMSE) in the course of training of model of type NNARX for different bases of neural network structures are resulted. Similar results have been received by the author for others extended the autoregressive predictors models NNARXMAX (NNARX + Moving Average, exogenous signal), NNOE (Neural Network Output Error).

In addition to these standard criteria, the authors in alternative works tested other indicators (for example, the ability to generalize, conditioning, statistical hypotheses, etc.). These results also give quite good indicators. [8]

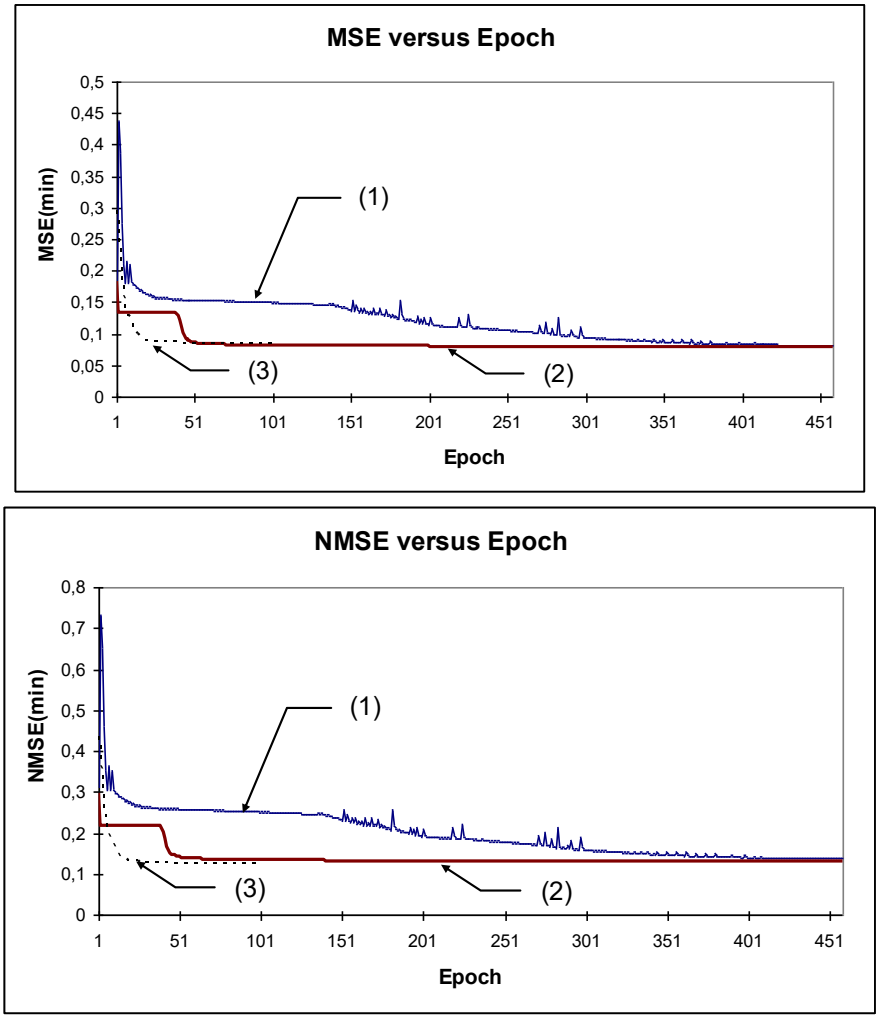

Figure 1: Change of criterions MSE and NMSE from quantity of iterations (epoch) at training neural identification model NNARX:
1 – two-layer perceptron which was trained for CG-method;
2 – a network of radial-basic functions (RBF) for GN-method;
3 – full coherent and partially recurrent a network for LM-method.

4. Conclusions

The analysis of results of computer modelling allows to make certain generalisations in the form of such **conclusions**.

Results of training intellectual neural models of type NNARX qualitatively almost identical if them accordingly to group (clusterization) behind identical methods of training (GN, CG, LM).

From the point of view of speed of convergence and robust the most perspective the method of Levenberg-Markuardt (LM), but it utilization of resources the greatest looks computing. The standard method of training of the NN, based on return distribution of an error (BP), has shown good enough robust, but its speed of coincidence slow enough, and requirements concerning resources – to big. Approximately identical and balanced enough results have shown methods of Gauss-Newton (GN) and Conjugate gradient (CG).

In view of the above-stated tests it is possible to recommend to apply to approximation difficult TP and using recurrent dynamic neural structure under condition of possibility of their hardware realisation (for example, neuro-graphic processors) or application of the parallel and distributed computing [9-11]. The latest is immediate prospects for continuation of the further researches in this direction.

5. References

[1] Kupin, A., Senko, A. Principles of intellectual control and classification optimization in conditions of technological processes of beneficiation complexes, CEUR Workshop Proceedings, 2015, 1356, pp. 153–160. URL: http://ceur-ws.org/Vol-1356/paper_34.pdf

[2] Bublikov, A.V., Tkachov, V.V. Automation of the control process of the mining machines based on fuzzy logic, Naukovyi Visnyk Natsionalnoho Hirnychoho Universytetu, 2019, 2019(3), pp. 112–118. DOI: 10.29202/nvngu/2019-3/19.
URL: https://www.metaljournal.com.ua/assets/Journal/11.2014.pdf

[3] Charu C.A. Neural Networks and Deep Learning, IBM T. J. Watson Research Center International Business MachinesYorktown HeightsUSA, Springer International Publishing AG, part of Springer Nature (2018). DOI: https://doi.org/10.1007/978-3-319-94463-0.

[4] Morkun, V.S., Morkun, N.V., Tron, V.V., Dotsenko, I.A. Adaptive control system for the magnetic separation process, Sustainable Development of Mountain Territories, 2018, 10(4), pp. 545–557. URL: http://naukagor.ru/Portals/4/%233%202018/№4,%202018.pdf?ver=2019-02-21-091240-697.

[5] Livshin I. Artificial Neural Networks with Java, Apress, Berkeley, CA (2019). DOI https://doi.org/10.1007/978-1-4842-4421-0

[6] Rudenko, O.G., Bezsonov, A.A. Neural network approximation of nonlinear noisy functions based on coevolutionary cooperative-competitive approach, Journal of Automation and Information Sciences, 2018, 50(5), pp. 11–21. DOI: 10.1615/JAutomatInfScien.v50.i5.20.

[7] Trunov, A., Malcheniuk, A. Recurrent Network As A Tool For Calibration In Automated Systems And Interactive Simulators, Eastern-European Journal of Enterprise Technologies, 2018, 2(9-92), pp. 54–60. DOI: https://doi.org/10.15587/1729-4061.2018.126498.

[8] Kupin, A. Research of properties of conditionality of task to optimization of processes of concentrating technology is on the basis of application of neural networks. Metallurgical and Mining Industry, 2014, 6(4), pp. 51–55.
URL: https://www.metaljournal.com.ua/assets/Journal/11.2014.pdf

[9] Hu, Z., Bodyanskiy, Y., Tyshchenko, O.K. Self-learning procedures for a kernel fuzzy clustering system, Advances in Intelligent Systems and Computing, 2019, 754, pp. 487–497. URL: https://link.springer.com/chapter/10.1007%2F978-3-319-91008-6_49.

[10] Derbentsev, V., Semerikov, S., Serdyuk, O., Solovieva, V., Soloviev, V. Recurrence based entropies for sustainability indices, E3S Web Conf. Volume 166, 2020. The International Conference on Sustainable Futures: Environmental, Technological, Social and Economic Matters (ICSF 2020), pp. 1-7. DOI: https://doi.org/10.1051/e3sconf/202016613031.

[11] Drozd, O., Kharchenko, V., Rucinski, A., Kochanski, T., Garbos, R., Maevsky, D. Development of Models in Resilient Computing, Proceedings of 10th IEEE International Conference on Dependable Systems, Services and Technologies (DESSERT'2019), Leeds, UK, June 5-7 2019, pp. 2 – 7. https://doi.org/10.1109/DESSERT.2019.8770035.