Proceedings of the 17th World Congress
The International Federation of Automatic Control
Seoul, Korea, July 6-11, 2008

Structure adaptation of multi-layer perceptron network for on-line system identification ⋆

Pavel Hering Miroslav Šimandl

Department of Cybernetics, Faculty of Applied Sciences, University of West Bohemia, Univerzitní 8, 306 14, Pilsen, Czech Republic,
(e-mail: phering@kky.zcu.cz); (e-mail: simandl@kky.zcu.cz)

Abstract: Identification of nonlinear systems by a neural network is treated. The paper deals with a design of a suitable neural network structure to approximate a nonlinear function of the identified system. Contrary to the recent algorithms, the proposed structure adaptation algorithm can be applied on-line during the identification process. The designed algorithm consist of a statistical test for making decision about suitability of an a priori chosen network and then either a growing or a pruning according to the size of the network is applied. The acceptance or rejection of the model is realized by application of the statistical cumulative sum test from the decision making field. The growing part of the algorithm repeatedly utilizes principle of the learning methodology for detecting faults in nonlinear dynamical systems for adding neurons to the hidden layer. Finally, the pruning algorithm is based on a measure of sensitivity of the model output error to the removing of the network connections. The properties of the proposed structure adaptation algorithm are illustrated in a numerical example.

1. INTRODUCTION

Multi-layer perceptron (MLP) networks are widely applied for modeling, control and fault detection of complex industrial systems [Witczak, 2006]. However, the crucial issue of the network design is answering the question - how to select a suitable network structure in order to be able to approximate system's nonlinearities with a desired accuracy. As Cybenko [1989] proved that a MLP network with one hidden layer and appropriate number of hidden neurons can approximate any continuous nonlinear function with arbitrary accuracy, the task can be restricted to finding appropriate number of hidden neurons. Unfortunately, it is not possible to use physical insight for determination of the number of hidden neurons and hence other more or less heuristic methods have been proposed. In general, two basic approaches, growing or pruning of the network, are used for construction of the network structure.

All the growing methods start from a minimal or an a priori set small network and add neurons until the desired accuracy of the model is achieved. Therefore, they prefer smaller networks having better generalization ability to the larger ones. The most famous growing method named cascade correlation method has been proposed by Fahlman and Lebiere [1990]. This method constructs the network with specific complex structure having multiple hidden layers and each neuron is connected directly to the output layer. Another method constructing a network with common layered feed-forward structure was developed by Ma and Khorasani [2003]. Survey of the main growing algorithms was published by e.g. Kwok and Yeung [1997].

All of the growing methods add neurons or layers iteratively by repeated utilization of the whole set of measurements, therefore they can not be used for on-line system identification as it is intended in the paper.

The second approach to network structure determination is based on pruning unnecessary connections or neurons from a larg network. In principle, there are two broad groups of the pruning methods [Reed, 1993]. The first one estimates the sensitivity of the error function to removal of an element and subsequently removes the connections or neurons with the least effect. The second one supplements a penalty term to an objective function, which causes that the unnecessary weights tend to zero value and are removed during the training. Between these two groups there is obviously an overlap since the objective function can have sensitivity terms.

During the years, close attention was paid to the design of the pruning methods and many of them were developed. Probably the most popular pruning methods are Optimal Brain Damage [LeCun et al., 1990] and the Optimal Brain Surgeon [Hassibi and Stork, 1993] which use the Hessian matrix or its approximation as a measure of saliency of the weights. Lauret et al. [2006] proposed another approach determining the neurons relevance from an analysis of the Fourier decomposition of the model output variance. However, these pruning methods working with the whole set of measurements are computationaly demanding and hence they can be used off-line only.

An approach usable for on-line neural network pruning was developed by Sum et al. [1999]. Contrary to the methods mentioned above, it treats the MLP network parameters as random variables and uses the Bayesian approach for the network training. However, the usage of Gaussian approximation of the parameters probability density function (pdf) causes the performance to be affected by the initial choice of the parameters. This problem is solved by the method designed by Šimandl and Hering [2005] considering more general pdf described by a mixture of Gaussian distributions [Sorenson and Alspach, 1971].

⋆ The work was supported by the Czech Science Foundation, projects GA 102/06/P202 and GA 102/08/0442.

978-1-1234-7890-2/08/$20.00 © 2008 IFAC 10.3182/20080706-5-KR-1001.2447

The joint drawback of all pruning methods is that it is difficult to decide how large an a priori network should be. Hence, some experience is required with the system which will be identified.

Due to the serious flaws of the present structure optimization methods, a new method for on-line structure adaptation will be designed in the paper. Particularly, the goal is to design an algorithm, which automatically decides about suitability of an a priori chosen model and adjusts its structure during the on-line identification of nonlinear stochastic system either by adding or removing neurons and weights. Thus, the proposed algorithm will consist of both - pruning and growing - parts.

The paper is organized as follows: In Section 2 the problem of identification of a nonlinear Gaussian stochastic system by the MLP network is formulated. Section 3 describes Gaussian sum filter applied in estimation of the conditional pdf's of the network parameters. The structure adapting algorithm is proposed in Section 4. In Section 5, an application of the proposed algorithm is illustrated in a numerical example. Finally, Section 6 summarizes the obtained results.

2. PROBLEM STATEMENT

Consider a nonlinear stochastic system with the single input u_k and the single output y_k, given by

$$y_k = \mathrm{h}(\varphi_k) + e_k, \tag{1}$$

$$\varphi_k = [y_{k-1}, \dots, y_{k-n_a}, u_{k-1}, \dots, u_{k-n_b}]^T, \tag{2}$$

where k is time instant, n_a and n_b are known constants, $\mathrm{h}(\varphi_k)$ is an unknown continuous function representing nonlinear dynamics of the system, φ_k denotes a vector of n_φ past inputs and outputs of the system, $\{e_k\}$ is zero mean Gaussian white noise with variance $E[e_k^2] = \sigma^2$.

The nonlinear function $\mathrm{h}(\varphi_k)$ is approximated by a two-layer perceptron network $\hat{\mathrm{h}}(\varphi_k, \mathbf{\Theta})$. The network consists of an a priori set of n_h neurons in the hidden layer and a single output linear neuron.

Mathematical description of the network is given by the following relations

$$\hat{y}_k = \hat{\mathrm{h}}(\varphi_k, \mathbf{\Theta}) = \sum_{j=1}^{n_h} W_j \cdot \tanh\left(\mathbf{w}_j^T \cdot \begin{bmatrix}\varphi_k \\ 1\end{bmatrix}\right) + W_0 \tag{3}$$

where $\hat{y}_k$ is output from the network at time instant k, $\mathbf{w}_j = [w_{j,0} \dots w_{j,n_\varphi,k}]^T$ and $\mathbf{W} = [W_0, \dots, W_{n_h}]^T$ denote vector of weights of inputs into the j^{th} hidden neuron and into the output neuron, respectively, $\mathbf{\Theta} = [\mathbf{w}_1^T, \dots, \mathbf{w}_{n_h}^T, \mathbf{W}^T]^T$. The vector of parameters $\mathbf{\Theta}$ is considered as a random variable described by a conditional pdf $p(\mathbf{\Theta}|\mathbf{y}^k)$, where $\mathbf{y}^k \triangleq \{y_0, y_1, \dots, y_k\}$ is a measurements history and $\mathbf{y}^{-1} \triangleq \varnothing$.

As the choice of the optimal number of hidden neurons n_h is impossible from the a priori information, the aim is to decide about suitability of the a priori network to approximate system's nonlinearities with a desired accuracy. Furthermore, an algorithm, which is capable to adjust the network structure with respect to the on-line measured data on the identified system, should be designed. Hence, the main goal is to design such algorithm which will have pruning and growing capabilities.

In order to be able to make any decisions about the network, its parameters need to be set. Therefore, some estimation method have to be applied at first. The next section deals with the discussion about an appropriate estimation method which could be used.

3. PARAMETERS ESTIMATION OF MLP NETWORK BY THE GAUSSIAN SUM METHOD

In the past, various approaches were developed for parameters estimation particularly based on minimization of prediction error, see Nørgaard et al. [2000], where the well-known back-propagation algorithm belongs, or on application of the Bayesian approach using local nonlinear estimation methods mainly represented by the extended Kalman filter (EKF) and the unscented Kalman filter, see Singhal and Wu [1988], van der Merwe and Wan [2001]. The increase of computational power in the recent years has enabled to use more powerful and computationally demanding global nonlinear estimation methods for the MLP network parameters estimation that are represented by GS method and particle filters, see Sorenson and Alspach [1971], Hering and Šimandl [2006], Neal [1996]. The global methods compute conditional probability density function (pdf) of the parameters thus they give the parameters estimates of higher quality and less influenced by the chosen initial conditions than the local ones. The main advantages of the GS method over the particle filters are a possibility to find an analytic solution for the pdf estimate and lower computational demands connected with comparable degree of accuracy. Therefore, this approach will be preferred for estimation of the pdf in the paper.

Let the a priori pdf of the parameters $\mathbf{\Theta}$ be assumed in the GS form

$$p(\mathbf{\Theta}|\mathbf{y}^{-1}) \triangleq \sum_{i=1}^{N} \alpha_{-1}^{(i)} \mathcal{N}\left\{\mathbf{\Theta} : \hat{\mathbf{\Theta}}_{-1}^{(i)}, \mathbf{P}_{-1}^{(i)}\right\}, \tag{4}$$

$$\sum_{i=1}^{N} \alpha_{-1}^{(i)} = 1, \quad \alpha_{-1}^{(i)} > 0,\ i = 1, \dots, N, \tag{5}$$

where N represents the number of terms in the mixture, $\mathcal{N}\left\{\mathbf{\Theta} : \hat{\mathbf{\Theta}}_{-1}^{(i)}, \mathbf{P}_{-1}^{(i)}\right\}$ denotes normal distribution of the random variable $\mathbf{\Theta}$ with mean $\hat{\mathbf{\Theta}}_{-1}^{(i)}$ and covariance matrix $\mathbf{P}_{-1}^{(i)}$.

The conditional pdf of the parameters at the time instant k is given by the recursive Bayesian relation

$$p(\mathbf{\Theta}|\mathbf{y}^k) = \frac{p(\mathbf{\Theta}|\mathbf{y}^{k-1})p(\mathbf{y}_k|\mathbf{\Theta})}{\int p(\mathbf{\Theta}|\mathbf{y}^{k-1})p(\mathbf{y}_k|\mathbf{\Theta})\mathrm{d}\mathbf{\Theta}}, \tag{6}$$

where $p(\mathbf{y}_k|\mathbf{\Theta})$ denotes the measurement pdf.

As the equation (3) is nonlinear, the function $\hat{\mathrm{h}}(\varphi_k, \mathbf{\Theta})$ is linearized at the points $\hat{\mathbf{\Theta}}_{k-1}^{(i)}$ to ensure the analytic solution of the Bayesian relation (6). For notational convenience, the variable φ_k is omitted below. Using the first order Taylor series expansion of the function $\hat{\mathrm{h}}(\mathbf{\Theta})$ having form

$$\hat{\mathrm{h}}(\mathbf{\Theta}) \approx \hat{\mathrm{h}}(\hat{\mathbf{\Theta}}_{k-1}^{(i)}) + \mathbf{H}_k^{(i)}(\hat{\mathbf{\Theta}}_{k-1}^{(i)})[\mathbf{\Theta} - \hat{\mathbf{\Theta}}_{k-1}^{(i)}],$$

where

$$\mathbf{H}_k^{(i)}(\hat{\mathbf{\Theta}}_{k-1}^{(i)}) = \mathbf{H}_k^{(i)} \triangleq \frac{\partial \hat{\mathrm{h}}(\mathbf{\Theta})}{\partial \mathbf{\Theta}}\Big|_{\mathbf{\Theta}=\hat{\mathbf{\Theta}}_{k-1}^{(i)}},$$

the conditional pdf of parameters $p(\mathbf{\Theta}|\mathbf{y}^k)$ is given as follows

$$p(\mathbf{\Theta}|\mathbf{y}^k) = \sum_{i=1}^{N} \alpha_k^{(i)} \mathcal{N}\left\{\mathbf{\Theta} : \hat{\mathbf{\Theta}}_k^{(i)}, \mathbf{P}_k^{(i)}\right\}, \tag{7}$$

where

$$\hat{\mathbf{\Theta}}_k^{(i)} = \hat{\mathbf{\Theta}}_{k-1}^{(i)} + \mathbf{K}_k^{(i)} \left[y_k - \hat{y}_k^{(i)} \right] \tag{8}$$

$$\mathbf{P}_k^{(i)} = \mathbf{P}_{k-1}^{(i)} - \mathbf{K}_k^{(i)} \mathbf{H}_k^{(i)} \mathbf{P}_{k-1}^{(i)} \tag{9}$$

$$\mathbf{K}_k^{(i)} = \mathbf{P}_{k-1}^{(i)} [\mathbf{H}_k^{(i)}]^T \left[\mathbf{H}_k^{(i)} \mathbf{P}_{k-1}^{(i)} [\mathbf{H}_k^{(i)}]^T + \sigma^2 \right]^{-1} \tag{10}$$

$$\alpha_k^{(i)} = \alpha_{k-1}^{(i)} \zeta_k^{(i)} / \sum_{i=1}^{N} \alpha_{k-1}^{(i)} \zeta_k^{(i)} \tag{11}$$

$$\zeta_k^{(i)} = \mathcal{N} \left\{ y_k : \hat{y}_k^{(i)}, \mathbf{H}_k^{(i)} \mathbf{P}_{k-1}^{(i)} [\mathbf{H}_k^{(i)}]^T + \sigma^2 \right\} \tag{12}$$

$$\hat{y}_k^{(i)} = \hat{\mathrm{h}}(\hat{\mathbf{\Theta}}_{k-1}^{(i)}) \tag{13}$$

for $i = 1, 2, \ldots, N$.

The relations (8) - (13) represent the GS method which is in fact a bank of N EKF's working in parallel. The multiple linearization of the function $\hat{\mathrm{h}}(\mathbf{\Theta})$ accomplished at the several points $\hat{\mathbf{\Theta}}_{k-1}^{(i)}$ improves performance and stability of the estimation algorithm.

The presented estimation algorithm introduces the general approach for the parameters estimation of the network being created during the structure adaptation.

4. STRUCTURE ADAPTATION

The a priori structure need not be optimal as it can be larger or smaller than it is necessary. Therefore, it is important to consider a structure adaptation and to find and apply a growing or pruning method to the current network.

4.1 Model verification

The problem of decision about model suitability has been extensively studied in fault detection area [Basseville and Nikiforov, 1993] where hypotheses testing methods are widely used for this purpose. Decisions can be carried out on the basis of the measurement obtained at the present time instant only or by using a measurements history. The snapshot and sequential statistical tests have been proposed for these purposes. Sequential tests overcome the snapshot tests in reliability due to the usage of more information. Therefore, some of them will be used for acceptance or rejection of selected neural network structure.

A simple sequential statistical method for testing the hypotheses, which can be applied, is Wald's sequential probability ratio test (SPRT) [Zhang, 1989, Basseville and Nikiforov, 1993]. The SPRT is commonly used for testing two alternative hypotheses.

Theoretically, the system function $\mathrm{h}(\cdot)$ could be approximated by the network with infinite number of hidden units exactly and the output error $\hat{e}_k$, defined as

$$\hat{e}_k = y_k - \sum_{i=1}^{N} \alpha_k^{(i)} \hat{y}_k^{(i)}, \tag{14}$$

would have the same features as the system noise but it is obviously not possible to achieve. In fact, the output error is affected by bias, estimated parameters values and the system noise [Cohn, 1996]. If the model has sufficiently large structure, the bias is negligible and could be omitted. If the parameters estimates are considered to be near the optimal values due to the usage of the global estimation method, then it is possible define the hypothesis ${}^0\mathcal{H}$ for acceptance of the given model that the output error has zero mean and some variance $\sigma_{\hat{e},0}^2$ corresponding to desired degree of the model accuracy. Nevertheless, this hypothesis could be kept also for larger networks than necessary. Hence, the pruning algorithm should be applied to find and remove unnecessary connections. Finally, the alternative hypothesis ${}^1\mathcal{H}$ can be defined as that the output error $\hat{e}_k$ has higher variance $\sigma_{\hat{e},1}^2$ than it is assumed by the hypothesis ${}^0\mathcal{H}$.

As the output errors $\hat{e}_k$ at individual time instants k are mutually independent, the conditional pdf $p(\hat{\mathrm{e}}_{t_0}^k|{}^j\mathcal{H})$ of the sequence of the errors $\hat{\mathrm{e}}_{t_0}^k$ from some initial time instant t_0 up to the current time k conditioned by validity of hypothesis ${}^j\mathcal{H}$ is given by product of the conditional pdf's $p(\hat{e}_t|{}^j\mathcal{H})$, $t = t_0, \ldots, k$, thus

$$p(\hat{\mathrm{e}}_{t_0}^k|{}^j\mathcal{H}) = \prod_{t=t_0}^{k} p(\hat{e}_t|{}^j\mathcal{H}), \tag{15}$$

where

$$p(\hat{e}_t|{}^j\mathcal{H}) = \mathcal{N}\{\hat{e}_t : 0, \sigma_{\hat{e},j}^2\}, \qquad j = 0, 1. \tag{16}$$

Let the sequential test statistics Λ_k based on the measurements from time t_0, $t_0 \leq k$, be given as a logarithm of a ratio of conditional probabilities (15), for $j = 0, 1$, thus

$$\Lambda_k = \ln \frac{p(\hat{\mathrm{e}}_{t_0}^k|{}^1\mathcal{H})}{p(\hat{\mathrm{e}}_{t_0}^k|{}^0\mathcal{H})}, \tag{17}$$

which can be rewritten to the recursive form

$$\Lambda_k = \Lambda_{k-1} + \lambda_k, \qquad \Lambda_{t_0-1} = 0, \tag{18}$$

$$\lambda_k = \ln \frac{p(\hat{e}_k|{}^1\mathcal{H})}{p(\hat{e}_k|{}^0\mathcal{H})}. \tag{19}$$

In order to decide about acceptance or rejection of the model, two constants $\gamma = P({}^1\mathcal{H}|{}^0\mathcal{H})$ and $\beta = P({}^0\mathcal{H}|{}^1\mathcal{H})$ corresponding to desired probabilities of rejection of hypothesis ${}^0\mathcal{H}$ when it holds and acceptance of ${}^0\mathcal{H}$ when ${}^1\mathcal{H}$ holds, respectively, have to be chosen a priori. They must satisfy the following condition

$$\gamma + \beta < 1 \tag{20}$$

Then, the decision about the model is made according to the following rules:

- $\Lambda_k \geq A$, then accept hypothesis ${}^1\mathcal{H}$,
- $\Lambda_k \leq B$, then accept hypothesis ${}^0\mathcal{H}$,
- $B < \Lambda_k \leq A$, then continue with next measurement without making any decision.

The constants A and B are chosen to meet the following inequalities

$$A \leq \ln \frac{1-\beta}{\gamma}, \tag{21}$$

$$B \geq \ln \frac{\beta}{1-\gamma}, \tag{22}$$

$$A > 0 > B. \tag{23}$$

It is reasonable to set the upper and lower bounds for value of Λ_k to reduce time delays in the decisions making that are commonly set three times as large as corresponding thresholds, i.e. $3A$ and $3B$ [Zhang, 1989].

The network with a given structure could approximate not only one function but a particular class of nonlinear functions. Therefore, the test could not be utilized starting from the initial time instant but after several training steps when the parameters values of the given network are adjusted to the concrete application and the convergency of the features of network output error has been reached. From experiments it arose that it is needed to use about $3n_\Theta$ measurements before beginning the model test.

4.2 Growing of neural network

The designed constructive approach is based on the principle of learning methodology to fault diagnosis in nonlinear dynamical systems. The main idea of this approach is to approximate any off-nominal behavior in the dynamical system by using on-line approximation structures and nonlinear estimation methods. When the failure arises, the on-line approximator is used to estimate it [Trunov and Polycarpou, 2000]. In this case the fault could represent inability of the a priori network to approximate system function with a desired accuracy.

When the current model is rejected by the statistical test at time k, the relation describing identified system is considered in the following form

$$y_k = \hat{\mathrm{h}}(\varphi_k, \mathbf{\Theta}) + \mathrm{f}({}^{\mathrm{f}}\mathbf{\Theta}, \varphi_k) + e_k, \tag{24}$$

where $\hat{\mathrm{h}}(\cdot)$ represents nonlinear behavior of the identified system, in this case it is represented by the a priori chosen neural network. Function $\mathrm{f}(\cdot)$ is the neural network with parameters ${}^{\mathrm{f}}\mathbf{\Theta}$ which is being used for modeling of possible faults in the system. For purposes of design of the constructive algorithm, it is used for modeling of the nonlinearities, which can not be expressed by the network $\hat{\mathrm{h}}(\cdot)$. The real behavior of the system is then approximated by the network $\hat{\mathrm{h}}(\cdot)$ together with $\mathrm{f}(\cdot)$.

However, design of the function $\mathrm{f}(\cdot)$ brings the same problem with determining the optimal network structure as the design of $\hat{\mathrm{h}}(\cdot)$. Hence, the minimal possible structure represented by network with one hidden neuron is selected for the purposes of the constructive algorithm, which is equivalent to addition of one new neuron into the hidden layer of the network $\hat{\mathrm{h}}(\cdot)$. Thus

$$\mathrm{f}({}^{\mathrm{f}}\mathbf{\Theta}, \varphi_k) = {}^{\mathrm{f}}W \cdot \tanh\left({}^{\mathrm{f}}\mathbf{w}^T \cdot \begin{bmatrix}\varphi_k \\ 1\end{bmatrix}\right), \tag{25}$$

where ${}^{\mathrm{f}}\mathbf{\Theta} = [{}^{\mathrm{f}}W, {}^{\mathrm{f}}\mathbf{w}^T]^T$ denotes $n_f = n_\varphi + 2$ parameters of the network $\mathrm{f}(\cdot)$.

Further, the parameters ${}^{\mathrm{f}}\mathbf{\Theta}$ have to be estimated. Let an a priori pdf of parameters ${}^{\mathrm{f}}\mathbf{\Theta}_k$ be like in (4) assumed in the GS form

$$p({}^{\mathrm{f}}\mathbf{\Theta}|\mathbf{y}^{-1}) \triangleq \sum_{i=1}^{N} {}^{\mathrm{f}}\alpha_{-1}^{(i)} \mathcal{N}\left\{{}^{\mathrm{f}}\mathbf{\Theta} : {}^{\mathrm{f}}\hat{\mathbf{\Theta}}_{-1}^{(i)}, {}^{\mathrm{f}}\mathbf{P}_{-1}^{(i)}\right\}, \tag{26}$$

$$\sum_{i=1}^{N} {}^{\mathrm{f}}\alpha_{-1}^{(i)} = 1, \quad {}^{\mathrm{f}}\alpha_{-1}^{(i)} > 0,\ i = 1, \ldots, N, \tag{27}$$

where terms of pdf (26) represent the so called candidate states for transition from the a priori network $\hat{\mathrm{h}}(\cdot)$ to the network augmented by one hidden neuron.

Estimated pdf of parameters (7) of the network $\hat{\mathrm{h}}(\cdot)$ could be also interpreted as N neural networks modeling the system, each trained by the EKF's from several different initial point estimates and working in parallel. During the parameters estimation one of the weights $\alpha_k^{(i)}$, $i = 1 \ldots, N$, typically converges to 1 and the rest to zero. It corresponds to selection of the most probable local model from a set of N models. As the parameters space has many equivalent minima, several weights can converge to the same value. Then the corresponding models are, in fact, the same and only one of them suffices to select. Hence, the most probable value at time k from parameters estimates $\hat{\mathbf{\Theta}}_k^{(i)}$ and relevant covariance matrix $\mathbf{P}_k^{(i)}$, $i = 1, \ldots, N$, are considered to produce the right model of the system behavior without fault and will be denoted as ${}^*\hat{\mathbf{\Theta}}_k$ and ${}^*\mathbf{P}_k$. The parameters pdf of the new model is then given as

$$p(\mathbf{\Theta}|\mathbf{y}^k) = \sum_{i=1}^{N} {}^{\mathrm{f}}\alpha_{-1}^{(i)} \mathcal{N}\left\{\mathbf{\Theta} : \begin{bmatrix}{}^{\mathrm{f}}\hat{\mathbf{\Theta}}_{-1}^{(i)} \\ {}^*\hat{\mathbf{\Theta}}_k\end{bmatrix}, \begin{bmatrix}{}^{\mathrm{f}}\mathbf{P}_{-1}^{(i)} & \mathbf{0} \\ \mathbf{0} & {}^*\mathbf{P}_k\end{bmatrix}\right\}. \tag{28}$$

When the new neuron is added, the estimation algorithm (8) - (13) continues with using pdf (28). After about $3n_f$ estimation steps the verification of the model can be realized again. The process of addition of neuron is repeated until the model does not attain the desired accuracy.

Due to the application of global estimation method, the proposed approach represents a special case of the multi-valued transition with candidate states having the same structure and differing in parameters values. Whereas, if the local estimation method is used it corresponds to the single-valued transition, [Kwok and Yeung, 1997].

The growing algorithm based on global description of neural networks parameters given by pdf (26) can be defined as follows:

Algorithm 1: *Growing algorithm*

Step 1 Select the most probable value at time k from parameters estimates $\hat{\mathbf{\Theta}}_k^{(i)}$ with relevant covariance matrix $\mathbf{P}_k^{(i)}$, $i = 1, \ldots, N$, and denote them as ${}^*\hat{\mathbf{\Theta}}_k$ and ${}^*\mathbf{P}_k$.

Step 2 Initialize a new neuron by setting ${}^{\mathrm{f}}\hat{\mathbf{\Theta}}_{-1}^{(i)}$, ${}^{\mathrm{f}}\mathbf{P}_{-1}^{(i)}$ and ${}^{\mathrm{f}}\alpha_{-1}^{(i)}$.

Step 3 Augment current parameters vector by the parameters of the new neuron $\mathbf{\Theta} = [{}^{\mathrm{f}}\mathbf{\Theta}^T,\ \mathbf{\Theta}^T]^T$ and use the pdf (28).

4.3 Pruning of neural network

The applied pruning method has been proposed and derived in [Šimandl and Hering, 2005], therefore in this section the algorithm will be only briefly introduced.

The method is based on computation of saliency of individual connections or subsets of them by determining sensitivity of the output error on their removing. The saliencies are estimated by using the conditional pdf of the network parameters (7) and connections having saliency lower than any chosen threshold T_0 are removed from the network. Removing itself is performed by setting the corresponding parameter value to zero.

Let a vector with the τ^{th} pruned parameter in the i^{th} term of the pdf (7) at time k be defined as

$$\hat{\mathbf{\Theta}}_{[\tau_i],k}^{(i)} = [\hat{\theta}_{1,k}^{(i)}, \ldots, \hat{\theta}_{\tau_i-1,k}^{(i)}, 0, \hat{\theta}_{\tau_i+1,k}^{(i)}, \ldots, \hat{\theta}_{n_\Theta,k}^{(i)}]^T,$$

where $\tau_i \in \{1, \ldots, n_\Theta\}$, $i = 1, \ldots, N$.

Moreover, let $\tau_{i_{[1,n]}}$ denote a set of $n \in \{1, \ldots, n_\Theta\}$ pointers τ_i to the parameters corresponding to the connections which should be pruned and the i^{th} parameters vector with n elements set to zero is denoted as $\hat{\mathbf{\Theta}}_{\tau_{i_{[1,n]}},k}$.

The connections saliency is measured by variable T [Šimandl and Hering, 2005] which is given

$$T = \sum_{i=1}^{N} \alpha_k^{(i)} T^{(i)}, \tag{29}$$

where

$$T^{(i)} = \alpha_k^{(i)} \sigma^2 (\hat{\mathbf{\Theta}}_k^{(i)} - \hat{\mathbf{\Theta}}_{\tau_{i_{[1,n]}},k}^{(i)})^T [\mathbf{P}_k^{(i)}]^{-1} (\hat{\mathbf{\Theta}}_k^{(i)} - \hat{\mathbf{\Theta}}_{\tau_{i_{[1,n]}},k}^{(i)}). \tag{30}$$

If the value T is less than or equal to the chosen threshold T_0, then the connections will be pruned. Since finding all possible combinations of the parameters that could be pruned is difficult, the strategy is based on pruning parameters from each term independently so that the increase of the error caused by the i^{th} term is $\alpha_k^{(i)} T^{(i)} \leq \alpha_k^{(i)} T_0$. This choice ensures that after all prunings at the time step it holds that $T \leq T_0$.

The designed pruning algorithm can be summarized in the following several basic steps:

Algorithm 2: *Pruning algorithm*

```
Step 1 Evaluate $V_{\tau_i} = [\mathbf{P}_k^{(i)}]^{-1}_{\tau_i,\tau_i} [\hat{\theta}^{(i)}_{\tau_i,k}]^2$ for all
    $\tau_i \in \{1, \ldots, n_\Theta\}$ denoting saliency of the parameters.
Step 2 Rearrange the indexes $\tau_i = \{\tau_i\}$ according to the
    ascending $V_{\tau_i}$.
Step 3 Evaluate $T^{(i)}$ for the sets $\tau_{i_{[1,n]}}$ for $n = 1 \ldots, n_\Theta$.
Step 4 Remove all parameters for which $T^{(i)} < T_0$.
```

4.4 Adaptive system identification algorithm

The individual parts of the structure adaptation algorithm were proposed above, i.e. parameters estimation, model verification and structure adaptation parts. Now, the algorithm will be summarized in the system identification framework for better transparency.

Algorithm 3: *Adaptive system identification algorithm*

```
Step 0 (Initialization)
    (1) Choose the structure of initial neural network, i.e. the
        number of hidden neurons $n_h$, the network inputs $\varphi_k$
        and the initial pdf (4).
    (2) Choose the parameters of CUSUM test, i.e. the probabilities $\gamma$, $\beta$ and the variances $\sigma^2_{\hat{e},0}$, $\sigma^2_{\hat{e},1}$. Set $t_0 \leftarrow 3n_\Theta$
        and $\Lambda_{t_0-1} = 0$.
    (3) Set the threshold $T_0$ for pruning algorithm.
Step 1 (Parameters estimation) Estimate the network parameters by application of relations (8) - (13).
Step 2 (Model verification and structure adaptation)
    If $k > 3n_\Theta$ and at least for $3n_f$ steps none neuron has been
    added.
    (1) Compute $\Lambda_k$ according to equation (17).
    (2) If $\Lambda_k \geq A$, then reject the current model and add one
        neuron using Algorithm 1. Set $t_0 \leftarrow k + 3n_f$.
    (3) If $\Lambda_k \leq B$, then accept the model and try to simplify the
        network by pruning Algorithm 2.
Step 3 $k \leftarrow k + 1$ and go to step 1.
```

The proposed structure adaptation algorithm provides an extension of the algorithm given in Šimandl and Hering [2005] where only pruning was considered.

5. NUMERICAL EXAMPLE

The proposed approach to adaptive identification of nonlinear system given by Algorithm 3 will be illustrated in the following numerical example.

Suppose a discrete-time nonlinear system with Gaussian disturbances described by the following equation

$$\begin{aligned} y_k = &- 1.5 \frac{y_{k-1} y_{k-2}}{1 + y_{k-1}^2 + y_{k-2}^2} + \\ &+ 0.35 \sin(y_{k-1} + y_{k-2}) + 1.2 u_{k-1} + e_k, \end{aligned}$$

where the input u_k is generated from the uniform distribution within interval $[-5; 5]$. The sequence of disturbances $\{e_k\}$ is a zero mean white noise with variance $\sigma^2 = 0.0025$.

The assumed system is modeled by the a priori chosen two-layer MLP network having 5 neurons with hyperbolic tangent activation functions in the hidden layer. The suitability of the network will be tested during the identification process and possibly the structure will be adjusted to achieve desired degree of accuracy of the final model.

The initial conditions for the network parameters are given as follows

$$p(\mathbf{\Theta} | \mathbf{y}^{-1}) = \sum_{i=1}^{N} \frac{1}{N} \mathcal{N} \left\{ \mathbf{\Theta} : \hat{\mathbf{\Theta}}_{-1}^{(i)}, \frac{10}{N} \cdot \mathbf{I} \right\},$$

where the initial means $\hat{\mathbf{\Theta}}_0^{(i)}$ are generated from the uniform distribution on the interval $(-1; 1)$.

The threshold used in the pruning algorithm is set to $T_0 = 10^{-5}$. The probabilities for the CUSUM test are chosen as follows: $\gamma = 0.01$ and $\beta = 0.01$. Variances defining the hypothesis ${}^0\mathcal{H}$ and ${}^1\mathcal{H}$ is $\sigma^2_{\hat{e},0} = \sigma^2$ and $\sigma^2_{\hat{e},1} = 9\sigma^2$, respectively.

The GS algorithm was tested for several numbers of terms $N = 1, \ldots, 4$. When $N = 1$ the proposed algorithm corresponds to using approximation of the parameters pdf by normal distribution and application of the EKF to the parameters estimation.

The criterion used for computation of quality of the model is chosen as the mean square error of one step ahead prediction of the system output for 1500 steps:

$$MSE = \frac{1}{1500} \sum_{k=1}^{1500} (y_k - \hat{y}_k)^2. \tag{31}$$

Experiments are repeated over 1000 trials and obtained averaged results are then summarized in Tab. 1. It is obvious, that the increasing number of terms of the GS improves the mean square error (MSE) of the final model. Furthermore higher number of terms N causes the decrease of number of needed neurons in the hidden layer and whole parameters with a improved quality of the model.

Table 1. Results of simulations for increasing number of terms N of GS averaged over 1000 trials. MSE - mean square error; N_{neu} - number of neurons at the end of the experiment; N_{par} - number of parameters at the end of the experiment.

N	1	2	3	4
mean(MSE)	0.2363	0.0977	0.0696	0.0526
var(MSE)	0.3036	0.0148	0.0083	0.0011
mean(N_{neu})	26	20	18	17
var(N_{neu})	236	101	72	54
mean(N_{par})	85	65	59	53
var(N_{par})	2805	948	601	448

One concrete development of the number of neurons and parameters together with results of the statistical test are depicted in Figure 1. The given results were obtained for four terms, $N = 4$, of GS but similar development would arise for different number of the terms.

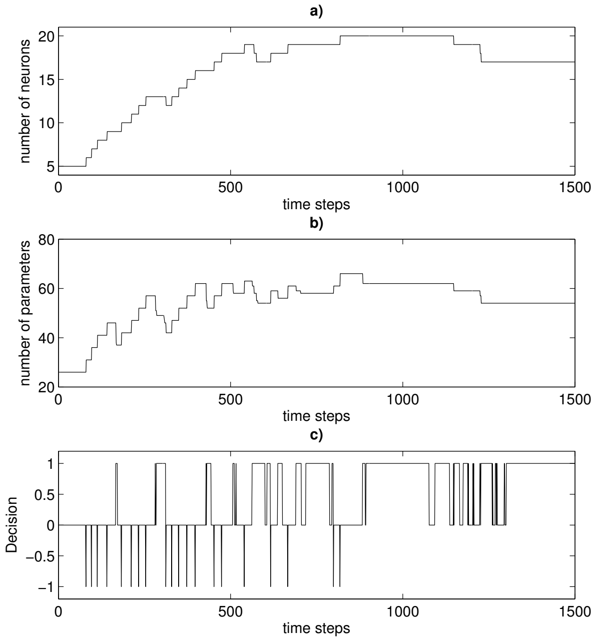

Fig. 1. Development of the neural network structure adaptation. **a)** Development of the number of neurons in the hidden layer; **b)** Development of the number of the network parameters; **c)** Decision about acceptance or rejection of current model. Acceptance of the model = 1; Rejection = -1; Continue without making decision = 0.

6. CONCLUSION

The new approach for on-line structure adaptation of the two-layer perceptron network in system identification of nonlinear stochastic systems was proposed. The proposed approach consists of the test of suitability of the a priori model and subsequent growing or pruning the network structure to obtain a desired model accuracy. The pdf of the network parameters is estimated by the global estimation method based on the Gaussian sum approach. Usage of this method brings the advantage of the multiple-valued transition to the network construction as it has the higher capability to adapt network to the problem at hand than by using a single valued-transition. If only one term in the sum is considered the single-valued transition is obtained. Increasing number of the terms in the estimated pdf helps to improve quality of the estimates of the network parameters and consequently to find a better network structure more independent of the choice of initial conditions. Obviously, it is connected with a linear increase of computational demands with respect to the number of the GS terms.

REFERENCES

M. Basseville and I. V. Nikiforov. *Detection of abrupt changes - Theory and application*. Prentice Hall, New Jersey, 1993.

D.A. Cohn. Neural network exploration using optimal experiment design. *Neural Networks*, 9(6):1071–1083, 1996.

G. Cybenko. Aproximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2: 304–314, 1989.

S. E. Fahlman and Ch. Lebiere. The cascade-correlation learning architecture. *Advances in Neural Information Processing Systems*, 2:524–532, 1990.

B. Hassibi and D. G. Stork. Second order derivatives for network pruning: Optimal brain surgeon. In *Advances in Neural Information Processing Systems*, volume 5, pages 164–171. Morgan Kaufmann, San Mateo, CA, 1993.

P. Hering and M. Šimandl. Gaussian sum based methods for neural network parameters estimation: aspects and comparison. In *Preprints of the 7th Portuguese Conference on Automatic Control CONTROLO'2006*, 2006.

Tin–Yau Kwok and Dit-Yan Yeung. Constructive algorithms for structure learning in feedforward neural networks for regression problems. *IEEE Transactions on Neural Networks*, 8(3):630–645, 1997.

P. Lauret, E. Fock, and T. A. Mara. A node pruning algorithm based on a fourier amplitude sensitivity test method. *IEEE Transactions on Neural Networks*, 17(2), 2006.

Y. LeCun, J. Denker, S. Solla, R. E. Howard, and L. D. Jackel. Optimal brain damage. In D. S. Touretzky, editor, *Advances in Neural Information Processing Systems II*, San Mateo, CA, 1990. Morgan Kauffman.

L. Ma and K. Khorasani. A new strategy for adaptively constructing multilayer feedforward neural networks. *Neurocomputing*, (51):361–385, 2003.

R. M. Neal. *Bayesian learning for neural networks, Lecture notes in statistics*. Springer-Verlag New York, 1996.

M. Nørgaard, O. Ravn, N.K. Poulsen, and L.K. Hansen. *Neural Networks for Modelling and Control of Dynamic Systems*. Springer-Verlag London, 2000. ISBN 1-85233-227-1.

R. Reed. Pruning algorithms - a survey. *IEEE Transactions on Neural Networks*, 4(5):740–747, 1993.

M. Šimandl and P. Hering. Recursive parameters estimation and structure adaptation of neural network. In *Proceedings of the 8th IASTED international conference on intelligent systems and control*, pages 78–83. Anaheim : ACTA Press, 2005.

S. Singhal and L. Wu. Training multilayer perceptrons with the extended Kalman algorithm. In *Advances in Neural Information Processing Systems*, volume 1, pages 133–140. Morgan Kaufmann Publishers, INC., 1988.

H.W. Sorenson and D.L. Alspach. Recursive Bayesian estimation using Gaussian sums. *Automatica*, 7:465–479, 1971.

J. Sum, Ch.-s. Leung, G.H. Young, and W.-k. Kan. On the Kalman filtering method in neural-network training and pruning. *IEEE Transactions on Neural Networks*, 10(1):161–166, 1999.

Alexander B. Trunov and Marios M. Polycarpou. Automated fault diagnosis in nonlinear multivariable systems using a learning methodology. *IEEE Transactions on Neural Networks*, 11(1):91–101, 2000.

R. van der Merwe and E.A. Wan. Efficient derivative-free Kalman filters for online learning. In *Proceedings of European Symposium on Artificial Neural Networks (ESANN)*, Bruges, Belgium, 2001.

M. Witczak. Toward training of feed-forward neural networks with the d-optimum input sequence. *IEEE Transactions on Neural Networks*, 17(2):357–373, 2006.

X.J. Zhang. *Auxiliary signal design in fault detection and diagnosis*. Springer-Verlag, 1989.