An Adaptive Forecasting of Nonlinear Nonstationary Time Series under Short Learning Samples

Elena Mantula[1] and Vladimir Mashtalir[1]

[1] Kharkiv National University of Radio Electronics, informatics department
Lenin ave., 14, 61166, Kharkiv, Ukraine

Mashtalir@kture.kharkov.ua, ElenaMantula@gmail.com

Abstract. Methods of nonstationary nonlinear time series forecasting under bounded a priori information provide an interdisciplinary applications area that is concerned with learning and adaptation of solutions from a traditional artificial intelligence point of view. It is extremely difficult to solve this type of problems in its general form, therefore, an approach based on the additive nonlinear auto regressive model with exogenous inputs and implemented on the base of parallel adalines set has been proposed. To find optimal combination of forecasts, an improvement of global random search has been suggested.

Keywords. Neural networks, forecasting model, combination of forecasts

Key terms. Environment, MathematicalModel

1 Introduction

'Conscious' decision making, in all possible varieties, is perhaps the most principal goal of artificial intelligence systems. Necessary 'creativity' implies the ability to produce novel solutions which are better than previous ones. The computational tools that assist in decision making should be such that they should take into all aspects of dissimilarity between a priori and a posteriori uncertainty. Uncertainty account is, per se, a manifestation of information deficiency, and relevant information is, on the contrary, a capacity to reduce uncertainty. An elimination of such rich in content gaps provides groundwork of knowledge engineering and management. In machine intelligence, manifold forecasts can be used for knowledge producing. The goal of the paper consists in reasonable (perfectly optimal) combination of forecasts to provide reliable semantic interpretation of achieved results with purpose knowledge generation.

Nowadays mathematical forecasting models of the behavior of objects, systems and phenomena in a wide variety of applications are well understood. There is a wealth of publications on this subject. It should be noted that the behavior of the objects is often given in the form of time series. Thus to forecast its behavior a variety of approaches to the analysis of time series can be used. Such approaches can be either traditional statistical methods (regression, correlation, spectral, Box-Jenkins) or adaptive, based on an exponential smoothing, tuning or learning forecasting models, or

intellectual, using vrious neural networks.

At present there are many objects (financial, economical, biomedical, etc.), described by time series containing unknown behavior trends, seasonal components, stochastic and random components, which significantly complicate synthesis of an effective predictive model. This complexity is especially pronounced in the environmental monitoring problems [1], where the analyzing time series have in equal measure stochastic and chaotic type of changes, have apparent nonstationarity and are subjected to striking changes.

In these conditions artificial ccc have proved to be useful tools in the best way [2-13]. As a rule, they realize so-called NARX-model [14], which has the form

$$\hat{y}(k) = f(y(k-1),...,y(k-n_{\mathrm{A}}), x(k-1),..., x(k-n_{\mathrm{B}}) \tag{1}$$

where $\hat{y}(k)$ is an estimation of forecasted variable $y(k)$ at discrete time $k = 1,2,...;$ $f(\circ)$ denotes certain nonlinear transform which is realized by a neural network; $x(k)$ is the observed exogenous factor that influences the behavior of $y(k)$; $n_{\mathrm{A}}, n_{\mathrm{B}}$ are observations memory parameters.

Moreover, it is not a matter of available observations insufficiency, since properties of time series (e.g. such indicator as air pollution in ecological forecasting) are changed so often that a neural network does not have time to detect separate stationary parts. In this connection there is a need to construct based on the neural network approach simplified predictive models for training which require the small enough volume data set.

2 Synthesis of a forecasting model

In conditions of input data lack instead of NARX-model (1) it is appropriate to use the so-called ANARX-model introduced in [15, 16] and fully investigated in [17, 18]. In general ANARX-model can be written as

$$\begin{aligned}\hat{y}(k) &= f_1(y(k-1), x(k-1)) + f_2(y(k-2), x(k-2)) + ... \\ &+ f_{max\{n_{\mathrm{A}}, n_{\mathrm{B}}\}}(y(k-n_{\mathrm{A}}), x(k-n_{\mathrm{B}})) = \\ &= \sum_{l=1}^{max\{n_A, n_B\}} f_l(y(k-l), x(k-l))\end{aligned} \tag{2}$$

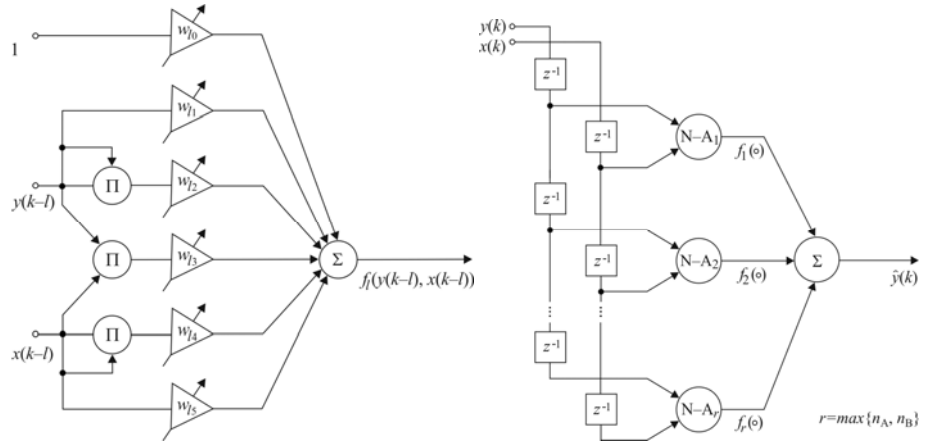
where original task is decomposed into many local ones with two input variables $y(k-l), x(k-l)$, $l = 1,2,...,max\{n_{\mathrm{A}},n_{\mathrm{B}}\}$.

For such nonlinear transforms it is quite convenient to use so-called N-adaline (abbr.: adaptive linear element) [19-21] that provide quadratic approximation of the data sequence. Fig. 1(a) demonstrates the architecture of N-adaline and (b) illustrates the architecture of ANARX-model constructed using N-adaline.

As we can see, N-adaline represents a generally accepted two-input adaline with a nonlinear preprocessor formed by three blocks of the product (Π) and the evaluator of the quadratic combination in the form

$$f_l(y(k-l), x(k-l)) = w_{l0} + w_{l1}y(k-l) + w_{l2}y^2(k-l) + w_{l3}y(k-l)x(k-l) + w_{l4}x^2(k-l) + w_{l5}x(k-l)$$

where each N-adaline contains 6 synaptic weights w_{lp}, $l = 1,2,...,max\{n_A, n_B\}$, $p = 0,1,...,5$. As a matter of fact, ANARX-model is formed by two lines of delay elements z^{-1} and $max\{n_A,n_B\}$ parallel learned N-adaline.



(a) N-adaline (b) ANARX models

Fig. 1. N-adaline and ANARX models based on N-adalines

Each from N-adalines is configured with any of the linear learning algorithms [22], however, it is clear that a limited amount of a priori information requires the use of time-optimal procedures. As such can be, for example, adaptive-multiplicative modification of Kachmarz adaptive algorithm [23], which assumes in this case the form

$$w_l(k) = w_l(k-1) + \gamma \frac{y(k) - w_l^{\mathrm{T}}(k-1)\varphi_l(k)}{\beta + \|\varphi_l(k)\|^2}\varphi_l(k) \qquad \textbf{(3)}$$

where $w_l = (w_{l_0}, w_{l_1}, w_{l_2}, w_{l_3}, w_{l_4}, w_{l_5})^{\mathrm{T}}$; $\varphi_l(k) = (1, y(k-l), y^2(k-l), y(k-l)x(k-l), x^2(k-l), x(k-l))^{\mathrm{T}}$; $0 < \gamma < 2$, $\beta \geq 0$ are some algorithm parameters selected on the base of empirical reasons.

If the data sequences are 'contaminated' by perturbations, instead of the one-step algorithm (3) it is profitably to apply procedures that provide filtering of perturbation and at the same time they have to be suitable for using in non-stationary conditions. It should be noted that modification of the recursive least squares method on a sliding window can be used [24]. The traditional estimation method of least squares on the window with s observations has the form

$$w_l(k) = \left(\sum_{\tau=k-s+1}^{k} \varphi_l(\tau)\varphi_l^{\mathrm{T}}(\tau)\right)^{-1} \sum_{\tau=k-s+1}^{k} \varphi_l(\tau) y(\tau)$$

and recurrent one can be presented as

$$\begin{cases} P(k) = P_s(k-1) - \dfrac{P_s(k-1)\varphi_l(k)\varphi_l^T(k)P_s(k-1)}{1+\varphi_l^T(k)P_s(k-1)\varphi_l(k)}, \\ P_s(k) = P(k) + \dfrac{P(k)\varphi_l(k-s)\varphi_l^{\mathrm{T}}(k-s)P(k)}{1-\varphi_l(k-s)P(k)\varphi_l(k-s)}, \\ p_s(k) = p_s(k-1) + \varphi_l(k)y(k) - \varphi_l(k-s)y(k-s), \\ w_l(k) = P_s(k)p_s(k). \end{cases} \quad \textbf{(4)}$$

We also note that if the algorithm (3) is in fact time-optimal gradient procedure, then the algorithm (4) is produced by Gaussian-Newton optimization procedure.

3 Optimal combination of forecasts

In real conditions the choice of the forecasting model structure is not a trivial task, especially that the same time series can be effectively described by a variety of different models. Also, the value of the lag orders $n_{\mathrm{A}}, n_{\mathrm{B}}$ remains unknown what makes it necessary to consider a set of competing models, and nonstationarity of analyzed series necessitates the use of various learning algorithms (in this case, (3), (4)) with different values γ,β,s. Thus, there arises a set of forecasts of the same process, from which we have to select the best.

To find the best forecast it is possible to use sufficiently effective approach, based on the optimal combination of forecasts [25], under which optimal in the sense of given criterion J^c linear combination is searching for a set of existing forecasts of the same series $\hat{y}_j(k),\ j = 1,2,...,m$

$$\hat{y}(k) = \sum_{j=1}^{m} c_j \hat{y}_j(k) \quad \textbf{(5)}$$

where the parameters of the combination satisfy the condition of unbiasedness

$$\sum_{j=1}^{m} c_j = 1. \quad \textbf{(6)}$$

In [25], an analytical approach to the weights c_j finding in (5) by optimizing the sum of squared errors criterion for forecasting with the constraints (6) is proposed. The use of one-step squared forecast errors criterion leads to the estimation

$$c_j(k) = \frac{\hat{y}_j(k)}{\sum_{j=1}^{m} \hat{y}_j(k)}.$$

However, combining of the analytical parameter estimates can be obtained under application of standard quadratic criterion J^c solely that specified by linearity of it derivatives so the solution of the problem reduces to solving a system of linear equations. At the same time for practitioners as a rule assess of the quality of forecasting

using the residual variance is unconvincing, and therefore characteristics allowing to estimate the accuracy in percentage are generally used, such as the criterion of a minimum of absolute percentage error

$$MAPE = \sum_{k=1}^{N} \left| \frac{y(k) - \hat{y}(k)}{y(k)} \right| 100\% \tag{7}$$

or maximum of the determination coefficient

$$R^2 = (1 - \frac{\sum_{k=1}^{N} (y(k) - \hat{y}(k))^2}{\sum_{k=1}^{N} (y(k) - \frac{1}{N} \sum_{k=1}^{N} y(k))^2}) 100\% . \tag{8}$$

It is obvious that in this case analytical estimations can not be obtained, and the use of gradient optimization procedures becomes more complicated due to sufficiently complex properties of functions (7), (8). In this connection the use of genetic algorithms is proposed in [26, 27]. Though such algorithms can find the global extremum, their own distinctive features are numerical awkwardness, they have a set of free parameters necessary defined by the user and at last it should be mentioned a low rate of convergence. Therefore, notice should be taken to more an efficient approach based on the random search [28] and its adaptive modifications. The most simple procedure, which allows to search for a global extremum, is walking random global search [28]. In general, this procedure is a statistic extension of the regular gradient search, and to provide the global search, random disturbance $\zeta(k)$ superimposes on character on a gradient movement what creates stochastic walking mode.

In the continuous case, the gradient method of minimization (maximization) of the goal function $J^c(t)$ is reduced to the motion of a point $c(t) = c_1(t),..., c_j(t),..., c_m(t)$ in m-dimensional space of adjustable parameters by a force directed toward the antigradient.

The trajectory of movement by antigradient $c(t)$ leads tuning process to a singular point. If starting point $c(0)$ belongs to an attraction region of global extremum then the corresponding trajectory will lead to a global minimum of the function $J^c(t)$. But if the point $c(0)$ does not belong to this region, the movement in the direction of antigradient will result in a local minimum, from which it is impossible to get out under the influence of forces directed by antigradient. Exactly because, it is helpful to use a random mechanism. Random shocks may help point $c(t)$ to overcome the barrier that separates the local minimum in which the learning process hit from the area in which the objective function $J^c(t)$ could further decrease. Under the influence of 'skew' toward anti-gradient and random shocks such movement is determined by the differential equation

$$\frac{dc(t)}{dt} = -\eta \nabla_c J^c(t) + \zeta(t)$$

where $\zeta(t)$ is m-dimensional normal random process with zero mathematical expec-

tation, delta-figurative autocorrelation function and components variance σ_ζ^2; η is parameter of step, ∇_c denotes gradient vector. It should be emphasized that for function (7) the components of the gradient can acquire the value $+1$ or -1. Generally, this algorithm provides searching for a global extremum [29].

Searching for global extremum can be speed up by reasonable selection of σ_ζ^2 and an adaptation during this process can be introduced in two ways. First, under introducing inertia in the learning process, it is possible to get a search similar to the movement by the method of 'heavy ball' [30]. Such movement is described by the differential equation

$$\frac{d^2c(t)}{dt^2} + b\frac{dc(t)}{dt} = -\eta\nabla J^c(t) + \zeta(t) \qquad \textbf{(9)}$$

where b is shockproofing coefficient (the more b, the less manifest of inserted inertia).

On time series processing, i.e. in discrete time, procedure (9) corresponds to the learning algorithm, described by the second order difference equation [31]

$$c(k) = c(k-1) + bc(k-2) - \eta(k)\nabla_c J^c(k) + \zeta(t) \qquad \textbf{(10)}$$

coinciding under $b = 0$ with walking random search. It is interesting to note that (10) is none other than the ARX- model of the second-order.

Second, the adaptation in the process of global search can be introduced by random process $\zeta(t)$ control, for example,

$$\frac{d\zeta(t)}{dt} = -\delta\zeta(t) - \eta_\zeta\frac{dJ^c(t)}{dt} + \sigma_\zeta^2 \mathrm{H}(t) \qquad \textbf{(11)}$$

where $\delta > 0$ is a autocorrelation parameter of random process $\zeta(t)$; $\mathrm{H}(t)$ is a vector of flat random noise. Introduce a modification of (11) in the discrete form

$$\zeta(k) = (1-\delta)\zeta(k-1) - \eta_\zeta(k)\Delta J^c(k) + \sigma_\zeta^2 \mathrm{H}(k) \qquad \textbf{(12)}$$

where Δ is the symbol of the first difference (discrete analogue of the derivative).

As it is easily seen from (11), (12), the optimization of the search process can be performed by appropriate selection of parameters δ, η_ζ and σ_ζ^2, since each of them acts on the certain properties of the search. Indeed, variation of the autocorrelation parameter δ determines the rate of the process $\zeta(k)$ decay that regulates its relations with the past. Thus, one can have an influence upon a search making it more or less dependent on the previous history if it is necessary.

Some few words of comment are desirable for parameters δ and η_ζ interaction explanation. If the search step η_ζ determines the intensity of accumulation of learning experience, then δ characterizes the level of this experience forgetting during the search. In this sense, these parameters are antagonistic. If in general $\delta = 0$ and there is no forgetting the vector $\eta_\zeta(k)$ increases in the direction of anti-gradient. Variance

of the process $\eta_\zeta(k)$ is determined by the value σ_ζ^2 and intensity of the flat random noise disturbance $\mathrm{H}(k)$. If σ_ζ^2 is sufficiently large then search may become unstable and, at low value, global properties are worsening. Thus, the use of a modified global random search allows simplify significantly the process of linear combination $c_j(k),\ j = 1,2,...,m$ tuning.

4 Conclusion

The problem of nonstationary nonlinear time series forecasting under bounded a priori information has been considered. An approach based on the additive nonlinear auto regressive model with exogenous inputs and implemented on the base of parallel adalines set has been proposed. To find optimal combination of forecasts, an improvement of global random search has been suggested. Distinctive feature of the approach is the computational simplicity and high performance attained by significant reducing the number of adjustable parameters.

References

1. Zanetti, P.: Air Pollution Modelling. Van Nostrand Reinhold, New York (1990)
2. Reich, S.L., Gomez, D.R., Dawidowski, L.E.: Artificial Neural Network for the Identification of Unknown Air Pollution Sources. Atmosphere Environment, Vol. 33, pp. 3045-3052 (1999)
3. Perez, P., Trier, A., Reyes, J.: Prediction of PM2.5 Concentration Several Hours in Advance Using Neural Networks in Santiago, Chile. Atmospheric Environmental, Vol. 34, pp. 1189–1196 (2000)
4. Niska, N., Hiltunen, T., Karppinen, A., Ruuskanen, J., Kolehmanen, M.: Evolving the Neural Network Model for Forecasting Air Pollution Time Series. Engineering Application of Artificial Intelligence, Vol. 17, 159–167 (2004)
5. Corani G.: Air Quality Prediction in Milan: Feed-Forward Neural Networks, Pruned Neural Networks and Lazy Learning. Ecological Modeling, Vol. 185, pp. 513–529 (2005)
6. Athanasiadis, I.N., Karatzas, K.D., Mitkas, P.A.: Classification Techniques for Air Quality Forecasting. In: Brewka G., Coradeschi S., Perini A. and Traverso P. (eds.): Proc. 17th European Conference on Artificial Intelligence, IOS Press, Amsterdam, 4.1–4.7 (2006)
7. Perez, P., Reyes, J.: An Integrated Neural Network Model for PM10 Forecasting. Atmospheric Environment. Vol. 40, pp. 2845–2857 (2006)
8. Lira, T.S., Barrozo, M.A.S., Assis. A.J.: Air Quality Prediction in Uberlandia, Brasil, Using Linear Models and Neural Networks. In: Plesu V., Agachi P. (eds.): Proc. 17th European Symp. on Computer Aided Process Engineering, Elsevier, Amsterdam, pp. 1–6 (2007)
9. Kurt, A., Gulbagci, B., Karaca, F., Alagha, O.: An Online Air Pollution Forecasting System Using Neural Networks. Environmental International, Vol. 34 (2008) 592–598
10. Carnevale, C., Finzi, G., Pisoni, E., Volta, M.: Neuro-Fuzzy and Neural Network Systems for Air Quality Control. Atmospheric Environmental, Vol. 43, pp. 4811–4821 (2009)

11. Nagendra, S.M., Shiva, Khare M.: Modelling Urban Air Quality Using Artificial Neural Network. Clean Technical Environmental Policy, Vol. 7, pp. 116–126 (2005)
12. Aktan, M, Bayraktar, H.: The Neural Network Modeling of Suspended Particulate Matter with Autoregressive Structure. Ekoloji, Vol. 19, No. 74, pp. 32–37 (2010)
13. Esau, I.: On Application of Artificial Neural Network Methods in Large-Eddy Simulations with Unresolved Urban Surfaces. Modern Applied Science, Vol. 4, No. 8, 3–11 (2010)
14. Nelles, O.: Nonlinear System Identification: From Classical Approaches to Neural Networks and Fuzzy Models. Springer, Berlin (2001)
15. Chowdhury, F.N., Input-Output Modeling of Nonlinear Systems with Time-Varying Linear Models. IEEE Trans. on Automatic Control, Vol. 45, No. 7, pp. 1355–1358 (2000)
16. Kotta, Ü., Sadegh, N.: Two Approaches for State Space Realization of NARMA Models: Bridging the Gap. Mathematical and Computer Modeling of Dynamical Systems, Vol. 8, No. 1, pp. 21–32 (2002)
17. Belikov, J., Vassiljeva, K., Petlenkov, E., Nomm S.: A Novel Taylor Series Based Approach for Control Computation in NN–ANARX Structure Based Control of Nonlinear Systems. In: Proc. 27th Chinese Control Conference, Beihang University Press, Kunming, pp. 474–478 (2008)
18. Vassiljeva, K., Petlenkov, E., Belikov, J.: State-Space Control of Nonlinear Systems Identified by ANARX and Neural Network Based SANARX Models. In: Proc. WCCI 2010 IEEE World Congress on Computational Intelligence, IEEE CSS, Piscataway, pp. 3816–3823 (2010)
19. Pham, D.T. Liu, X.: Modeling and Prediction Using GMDH Networks of Adalines with Nonlinear Preprocessors. Int. J. System Science Vol. 25, No. 11 (1994) 1743–1759
20. Pham, D.T. Liu, X.: Neural Networks for Identification, Prediction and Control. Springer, London (1995)
21. Rudenko, O.G., Bodyanskiy, Ie.V.: Artificial Neural Networks. SMIT, Kharkov (2005) (in Russian)
22. Haykyn, S.: Neural Networks: A Comprehensive Foundation. Prentice Hall. Inc., New York (1999)
23. Raybman, N.S., Chadeev, V.M.: Creating of Manufacture Process Models. Energiya, Moscow (1975) (in Russian)
24. Perelman, I.I.: Operative Identification of Control Objects. Energoatomizdat, Moscow (1982) (in Russian)
25. Sharkeya, A.J.C.: On Combining Artificial Neural Nets. Connection Science, Vol. 8, No. 3, pp. 299–314 (1996)
26. Zagoryjko, N.G.: Empirical Prediction. Nauka, Novosibirsk (1979) (in Russian)
27. Zagoryjko, N.G.: Applied Approach of Data Analysis, p. 264 (1999) (in Russian)
28. Rastrigin, L.A.: Statistical Search Technology. Nauka, Moscow (1968) (in Russian)
29. Rastrigin, L.A.: Systems of Extremal Control. Nauka, Moscow (1974) (in Russian)
30. Polyak B.T.: Introduction into Optimization. Nauka, Moscow (1983) (in Russian)
31. Bodyanskiy, Ie. V., Rudenko, O.G.: Artificial Neural Networks: Arhitectures, Learning, Applications. TELETEX, Kharkov (2004) (in Russian)